Accelerating Variance-Reduced Stochastic Gradient Methods

Derek Driggs*[1], Matthias J. Ehrhardt†[2] and Carola-Bibiane Schönlieb‡[1]

[1]Department of Applied Mathematics and Theoretical Physics, Cambridge University
[2]Institute for Mathematical Innovation, University of Bath

October 22, 2019

Abstract

Variance reduction is a crucial tool for improving the slow convergence of stochastic gradient descent. Only a few variance-reduced methods, however, have yet been shown to directly benefit from Nesterov's acceleration techniques to match the convergence rates of accelerated gradient methods. Such approaches rely on "negative momentum", a technique for further variance reduction that is generally specific to the SVRG gradient estimator. In this work, we show for the first time that negative momentum is unnecessary for acceleration and develop a universal acceleration framework that allows all popular variance-reduced methods to achieve accelerated convergence rates. The constants appearing in these rates, including their dependence on the dimension n, scale with the mean-squared-error and bias of the gradient estimator. In a series of numerical experiments, we demonstrate that versions of SAGA, SVRG, SARAH, and SARGE using our framework significantly outperform non-accelerated versions and compare favourably with algorithms using negative momentum.

1 Introduction

We are interested in solving the following composite convex minimisation problem:

$$\min_{x\in\mathbb{R}^m}\left\{F(x)\stackrel{\text{def}}{=}f(x)+g(x)\stackrel{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}f_i(x)+g(x)\right\}. \tag{1}$$

Throughout, we assume $f_i:\mathbb{R}^m\to\mathbb{R}$ are convex and have L-Lipschitz continuous gradients for all i. We also assume $g:\mathbb{R}^m\to\mathbb{R}\cup\{\infty\}$ is proper, lower semicontinuous, and μ-strongly convex with $\mu\geq 0$, but we do not require g to be differentiable. Problems of this form are ubiquitous in many fields, including machine learning, compressed sensing, and image processing (see, e.g., [9, 10, 21, 31]). Fundamental examples include LASSO [31] and matrix completion [10], where f is a least-squares loss and g is the ℓ_1 or nuclear norm, respectively, and sparse logistic regression, where f is the logistic loss and g is the ℓ_1 norm.

One well-studied algorithm that solves (1) is the *forward-backward splitting algorithm* [11, 25]. This method has a worst-case convergence rate of $\mathcal{O}(1/T)$ when F is not strongly convex, and when F is μ-strongly convex, it converges linearly with a rate of $\mathcal{O}\left((1+\kappa^{-1})^{-T}\right)$, where $\kappa\stackrel{\text{def}}{=}L/\mu$ is the condition number of F. The *inertial* forward-backward splitting algorithm [7] converges at an even faster rate of $\mathcal{O}\left(1/T^2\right)$ without strong convexity and a linear rate of $\mathcal{O}\left((1+\kappa^{-1/2})^{-T}\right)$ when F is strongly convex. The inertial forward-backward method is able to achieve these optimal convergence rates because it incorporates *momentum*, using information from previous iterates to adjust the current iterate.

Although the inertial forward-backward algorithm converges quickly, it requires access to the full gradient ∇f at each iteration, which can be costly, for instance, when n is large. In many applications, common

*d.driggs@damtp.cam.ac.uk
†m.ehrhardt@bath.ac.uk
‡cbs31@cam.ac.uk

problem sizes are so large that computing ∇f is prohibitively expensive. Stochastic gradient methods exploit the separable structure of f, using the gradient of a few of the components ∇f_i to estimate the full gradient at the current iterate. In most cases, the complexity of computing ∇f_i for one i is $1/n$-times the complexity of computing the full gradient, so stochastic gradient methods generally have a much smaller per-iteration complexity than full-gradient methods. Moreover, it has recently been shown that the optimal convergence rates of stochastic gradient methods are $\mathcal{O}\left(\sqrt{n}/T^2\right)$ without strong convexity and $\mathcal{O}\left(\theta_S^{-T}\right)$ with $\theta_S \stackrel{\text{def}}{=} 1 + \sqrt{\frac{\mu}{Ln}}$ when g is μ-strongly convex, matching the optimal dependence on T and κ of full-gradient methods [33].[1] Stochastic gradient methods have undergone several revolutions to improve their convergence rates before achieving this lower bound. We summarise these revolutions below, beginning with traditional stochastic gradient descent.

Stochastic Gradient Descent (SGD). *Stochastic gradient descent*, dating back to [28], uses the gradients $\nabla f_j,\ \forall j \in J_k \subset \{1, 2, \ldots, n\}$ to estimate the full gradient. The *mini-batch* J_k is an index set chosen uniformly at random from all subsets of $\{1, 2, \ldots, n\}$ with cardinality $b \stackrel{\text{def}}{=} |J_k|$. When $b \ll n$, the per-iteration complexity of stochastic gradient descent is much less than full-gradient methods. However, the per-iteration savings come at the cost of a slower convergence rate, as SGD converges at a rate of $\mathcal{O}(1/\sqrt{T})$ in the worst case. Still, SGD outperforms full-gradient methods on many problems, especially if a low-accuracy solution is acceptable.

Variance Reduction. Variance-reduced estimators use gradient information from previous iterates to construct a better estimate of the gradient at the current step, ensuring that the mean-squared error of these estimates decreases as the iterations increase. Variance-reduction improves the convergence rates of stochastic gradient methods, but either have a higher per-iteration complexity or have larger storage requirements than SGD. The two most popular variance-reduced algorithms are SVRG [18] and SAGA [13], which use the following estimators to approximate $\nabla f(x_{k+1})$:

$$\widetilde{\nabla}_{k+1}^{\text{SVRG}} \stackrel{\text{def}}{=} \frac{1}{b}\left(\sum_{j \in J_k} \nabla f_j(x_{k+1}) - \nabla f_j(\widetilde{x})\right) + \nabla f(\widetilde{x}) \tag{2}$$

$$\widetilde{\nabla}_{k+1}^{\text{SAGA}} \stackrel{\text{def}}{=} \frac{1}{b}\left(\sum_{j \in J_k} \nabla f_j(x_{k+1}) - \nabla f_j(\varphi_k^j)\right) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\varphi_k^i). \tag{3}$$

In SVRG, the full gradient $\nabla f(\widetilde{x})$ is computed every $m \approx 2n$ iterations, and $\nabla f(\widetilde{x})$ is stored and used for future gradient estimators. SAGA takes a similar approach, storing n past stochastic gradients, and updating the stored gradients so that $\nabla f_j(\varphi_{k+1}^j) = \nabla f_j(x_{k+1})$. In this work, we consider a variant of SVRG where the full gradient is computed at every iteration with probability $1/p \in (0, 1]$ rather than deterministically computing the full gradient every $2n$ iterations.

SVRG, SAGA, and related variance-reduced methods converge at a rate of $\mathcal{O}(n/T)$ when no strong convexity is present. With strong convexity, these algorithms enjoy linear convergence, with a rate of $\mathcal{O}(\left(1 + (n + \kappa)^{-1}\right)^{-T})$. Both of these rates match the rates of full-gradient methods in terms of their dependence on T and κ. Although these convergence rates are significantly faster than the rate of SGD, they do not achieve the fastest possible dependencies of $\mathcal{O}\left(1/T^2\right)$ without strong convexity and $\mathcal{O}((1+\kappa^{-1/2})^{-T})$ with strong convexity.

Variance Reduction with Bias. SAGA and SVRG are *unbiased* gradient estimators because they satisfy $\mathbb{E}_k \widetilde{\nabla}_{k+1} = \nabla f(x_{k+1})$, where $\mathbb{E}_k$ is the expectation conditioned on the first k iterates. There are several popular variance-reduced algorithms that use biased gradient estimators [23, 29]. In [14], the authors develop a framework for proving convergence guarantees for biased methods, suggesting that the convergence rates

[1] The results in [33] are complexity bounds, bounding the number of gradient and prox oracle calls required to achieve a given tolerance. For algorithms performing $\mathcal{O}(1)$ oracle calls per iteration, these complexity bounds imply the stated bounds on convergence rates.

of biased stochastic gradient estimators depend on the sum of two terms:

$$\gamma^2 \mathbb{E}_k \|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 + \gamma \left\langle \nabla f(x_{k+1}) - \mathbb{E}_k \widetilde{\nabla}_{k+1}, x_{k+1} - x^* \right\rangle.$$

These terms are the mean-squared error (MSE) of the gradient estimator and the "bias term", respectively. The authors also show that *recursive* gradient estimators such as SARAH [23] and SARGE [14] minimise these terms better than other biased or unbiased estimators, leading to better convergence rates in some settings. The SARAH gradient estimator is

$$\widetilde{\nabla}^{\text{SARAH}}_{k+1} \stackrel{\text{def}}{=} \begin{cases} \frac{1}{b}\left(\sum_{j \in J_k} \nabla f_j(x_{k+1}) - \nabla f_j(x_k)\right) + \widetilde{\nabla}^{\text{SARAH}}_k & \text{w.p. } 1 - \frac{1}{p}, \\ \nabla f(x_{k+1}) & \text{w.p. } \frac{1}{p}. \end{cases} \tag{4}$$

As with SVRG, we consider a slight variant of the SARAH estimator in this work, where we compute the full gradient at every step with probability $1/p$. The SARGE gradient estimator is similar to the SAGA estimator.

$$\widetilde{\nabla}^{\text{SARGE}}_{k+1} \stackrel{\text{def}}{=} \frac{1}{b}\left(\sum_{j \in J_k} \nabla f_j(x_{k+1}) - \psi^j_k\right) + \frac{1}{n}\sum_{i=1}^n \psi^i_k - \left(1 - \frac{b}{n}\right)\left(\frac{1}{b}\sum_{j \in J_k} \nabla f_j(x_k) - \widetilde{\nabla}^{\text{SARGE}}_k\right), \tag{5}$$

where the variables ψ^i_k follow the update rule $\psi^j_{k+1} = \nabla f_j(x_k) - \left(1 - \frac{b}{n}\right) \nabla f_j(x_{k-1})$ for all $j \in J_k$, and $\psi^i_{k+1} = \psi^i_k$ otherwise. Like SAGA, SARGE uses stored gradient information to avoid having to compute the full gradient. These estimators differ from SAGA and SVRG because they are biased (i.e., $\mathbb{E}_k \widetilde{\nabla}_{k+1} \neq \nabla f(x_{k+1})$). Many works have recently shown that algorithms using the SARAH or SARGE gradient estimators achieve faster convergence rates than algorithms using other estimators in certain settings. Importantly, these recursive gradient methods produce algorithms that achieve the *oracle complexity lower bound* for non-convex composite optimisation [14, 15, 26, 32, 38]. They have not yet been shown to achieve optimal convergence rates for convex problems.

Variance Reduction with Negative Momentum. Starting with Katyusha [2] and followed by many others [1, 3, 4, 19, 30, 36, 37], a family of stochastic gradient algorithms have recently emerged that achieve the optimal convergence rates implied by [33]. There are two components to these algorithms that make this acceleration possible. First, these algorithms incorporate momentum into each iteration, either through linear coupling [6], as in the case of [1, 2, 3, 4, 36], or in a more traditional manner reminiscent of Nesterov's accelerated gradient descent [30, 37]. Second, these algorithms incorporate an "anchor-point" into their momentum updates that supposedly softens the negative effects of bad gradient evaluations. Almost all of these algorithms are an accelerated form of SVRG with updates of the form

$$\begin{aligned} x_{k+1} &= \widetilde{x} + \tau_k(x_k - \widetilde{x}), \quad \text{or} \\ x_{k+1} &= \tau_1 z_k + \tau_2 \widetilde{x} + (1 - \tau_1 - \tau_2) y_k, \end{aligned}$$

using traditional acceleration or linear coupling, respectively (z_k and y_k are as defined in Algorithm 1, and $\tau_k, \tau_1, \tau_2 \in [0, 1]$). We see that these updates "attract" the current iterate toward a "safe" point, $\widetilde{x}$, where we know the full gradient. Because of this "attractive" rather than "repulsive" quality, updates of this type have been termed "negative momentum".

There are several issues with negative momentum. Most importantly, negative momentum is algorithm-specific. Unlike Nesterov's method of momentum or linear coupling, negative momentum cannot be applied to other stochastic gradient algorithms. SAGA, for example, cannot be accelerated using negative momentum of this form, because there does not exist a point $\widetilde{x}$ where we compute the full gradient (however, see [36]). Also, numerical experiments show that negative momentum is often unnecessary to achieve acceleration (see the discussion in [2] or Section 7, for example), suggesting that it is only a theoretical convenience for proving convergence rates.

Algorithm 1 A Universal Framework for Acceleration

Input: Set step size γ_k and momentum parameter τ_k as in Theorem 5 if $\mu = 0$ or as in Theorem 6 otherwise, and gradient estimator $\widetilde{\nabla}$.

Initialise $z_0 = y_0 = x_0$.
for $k = 0, 1, \cdots, T-1$ **do**
  $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k) y_k$.
  Compute $\widetilde{\nabla}_{k+1}$, an estimate of $\nabla f(x_{k+1})$.
  $z_{k+1} \leftarrow \text{prox}_{\gamma_k g}\left(z_k - \gamma_k \widetilde{\nabla}_{k+1}\right)$.
  $y_{k+1} \leftarrow \tau_k z_{k+1} + (1-\tau_k) y_k$.
end for

Contributions. In this work, we show that stochastic gradient algorithms do not require negative momentum to achieve accelerated convergence rates. We introduce the *MSEB property*, a property that implies natural bounds on the bias and MSE of a gradient estimator, and we prove accelerated convergence rates for all MSEB gradient estimators. As special cases, we show that incorporating the SAGA, SVRG, SARAH, and SARGE gradient estimators into the framework of Algorithm 1 creates a stochastic gradient method with an $\mathcal{O}\left(1/T^2\right)$ convergence rate without strong convexity, and a linear convergence rate that scales with $\sqrt{\kappa}$ when strong convexity is present, achieving the optimal convergence rates in both cases up to a constant depending on the bias and MSE of the estimator.

Roadmap. We introduce our algorithm and state our main result in Section 2. We compare our results to existing work in Section 3. The next four sections are devoted to proving our main results. In Section 4, we review elementary results on the subdifferential relation, results on the proximal operator, and lemmas from convex analysis. We prove a general inequality for accelerated stochastic gradient methods using any stochastic gradient estimator in Section 5. This inequality implies that many stochastic gradient methods can be accelerated using our momentum scheme; to prove an accelerated convergence rate for a specific algorithm, we only need to apply an algorithm-specific bound on the MSE and bias of the gradient estimator. We do this for the SAGA, SVRG, SARAH, and SARGE gradient estimators in Section 6. Finally, in Section 7, we demonstrate the performance of our algorithms in numerical experiments.

2 Algorithm and Main Results

The algorithm we propose is outlined in Algorithm 1. Algorithm 1 takes as input any stochastic gradient estimator $\widetilde{\nabla}_{k+1}$, so it can be interpreted as a framework for accelerating existing stochastic gradient methods. This algorithm incorporates momentum through linear coupling [6], but is related to Nesterov's accelerated gradient method after rewriting x_{k+1} as follows:

$$x_{k+1} = y_k + (1-\tau_k)(y_k - y_{k-1}).$$

With $\tau_k = 1$, there is no momentum, and the momentum becomes more aggressive for smaller τ_k.

We show that as long as the MSE and bias of a stochastic gradient estimator satisfies certain bounds and the parameters γ_k and τ_k are chosen correctly, Algorithm 1 converges at an accelerated rate. There are three principles for choosing γ_k and τ_k so that Algorithm 1 achieves acceleration.

1. The step size γ_k should be small, roughly $\mathcal{O}(1/n)$ with the exact dependence on n decreasing with larger MSE and bias of the gradient estimator.

2. On non-strongly convex objectives, the step size should grow sufficiently slowly, so that $\gamma_k^2(1-\rho) \leq \gamma_{k-1}^2\left(1-\frac{\rho}{2}\right)$ with $\rho = \mathcal{O}(1/n)$ decreasing with larger MSE and bias.

3. The momentum should become more aggressive with smaller step sizes, with $\tau_k = \mathcal{O}\left(\frac{1}{n\gamma_k}\right)$.

For strongly convex objectives, γ_k and τ_k can be kept constant.

For Algorithm 1 to converge, the stochastic gradient estimator must have controlled bias and MSE. Specifically, we require the estimator to satisfy the MSEB property,[2] introduced below.

Definition 1 *For any sequence $\{x_{k+1}\}$, let $\widetilde{\nabla}_{k+1}$ be a stochastic gradient estimator generated from the points $\{x_{\ell+1}\}_{\ell=0}^{k}$. The estimator $\widetilde{\nabla}_{k+1}$ satisfies the* MSEB$(M_1, M_2, \rho_M, \rho_B, \rho_F)$ *property if there exist constants $M_1, M_2 \geq 0$, $\rho_M, \rho_B, \rho_F \in (0, 1]$, and sequences $\mathcal{M}_k$ and $\mathcal{F}_k$ satisfying*

$$\nabla f(x_{k+1}) - \mathbb{E}_k \widetilde{\nabla}_{k+1} = (1 - \rho_B)\left(\nabla f(x_k) - \widetilde{\nabla}_k\right),$$

$$\mathbb{E}\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 \leq \mathcal{M}_k,$$

$$\mathcal{M}_k \leq \frac{M_1}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|^2 + \mathcal{F}_k + (1 - \rho_M)\mathcal{M}_{k-1},$$

and

$$\mathcal{F}_k \leq \sum_{\ell=0}^{k}\frac{M_2(1-\rho_F)^{k-\ell}}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{\ell+1}) - \nabla f_i(x_\ell)\|^2.$$

On a high-level, the MSEB property guarantees that the bias and MSE of the gradient estimator decrease sufficiently quickly with k.

Remark 1 *In [8], the authors study the convergence of unbiased stochastic gradient methods under first- and second-moment bounds on the gradient estimator. The bounds implied by the MSEB property are similar, but with the crucial difference that they are* non-Markovian*; we allow our bound on $\mathcal{M}_k$ to depend on all preceding iterates, not just x_k.*

In this work, we show that most existing stochastic gradient estimators satisfy the MSEB property, including SAGA, SVRG, SARAH, SARGE, and the full gradient estimator. We list their associated parameters in the following propositions.

Proposition 1 *The full gradient estimator $\widetilde{\nabla}_{k+1} = \nabla f(x_{k+1})$ satisfies the MSEB property with $M_1 = M_2 = 0$ and $\rho_M = \rho_B = \rho_F = 1$.*

Proof. The bias and MSE of the full gradient estimator are zero, so it is clear these parameter choices satisfy the bounds in the MSEB property. □

Although trivial, Proposition 1 allows us to show that our analysis recovers the accelerated convergence rates of the inertial forward-backward algorithm as a special case. The MSEB property applies to the SAGA and SVRG estimators non-trivially.

Proposition 2 *The SAGA gradient estimator* (3) *satisfies the MSEB property with $M_1 = \mathcal{O}(n/b^2)$, $\rho_M = \mathcal{O}(b/n)$, $M_2 = 0$, and $\rho_B = \rho_F = 1$. Setting $p = \mathcal{O}(n/b)$, the SVRG gradient estimator* (2) *satisfies the MSEB property with the same parameters.*

We prove Proposition 2 in Appendix B. We are able to choose $\rho_B = 1$ for the SAGA and SVRG gradient estimators because they are unbiased, and we can choose $M_2 = 0$ and $\rho_F = 1$ for these estimators because they admit Markovian bounds on their variance. This is not true for SARAH and SARGE, but these estimators are still compatible with our framework. We prove Propositions 3 and 4 in Appendices C and D, respectively.

Proposition 3 *Setting $p = \mathcal{O}(n)$, the SARAH gradient estimator* (4) *satisfies the MSEB property with $M_1 = \mathcal{O}(1/b)$, $M_2 = 0$, $\rho_M = \mathcal{O}(1/n)$, $\rho_B = \mathcal{O}(1/n)$, and $\rho_F = 1$.*

[2] Because this property asserts bounds on the mean-squared-error and bias of a stochastic gradient estimator, the name MSEB is a natural choice. We suggest the pronunciation "M-SEB".

Proposition 4 *The SARGE gradient estimator* (5) *satisfies the MSEB property with* $M_1 = \mathcal{O}(1/(bn))$, $M_2 = \mathcal{O}(1/n^2)$, $\rho_M = \mathcal{O}(b/n)$, $\rho_B = \mathcal{O}(b/n)$, *and* $\rho_F = \mathcal{O}(b/n)$.

All gradient estimators satisfying the MSEB property can be accelerated using the framework of Algorithm 1, as the following two theorems guarantee.

Theorem 5 (**Acceleration Without Strong Convexity**) *Suppose the stochastic gradient estimator* $\widetilde{\nabla}_{k+1}$ *satisfies the* MSEB$(M_1, M_2, \rho_M, \rho_B, \rho_F)$ *property. Define the constants*

$$\Theta_1 \overset{def}{=} 1 + \frac{8(1-\rho_B)}{\rho_B^2 \rho_M}, \quad \Theta_2 \overset{def}{=} \frac{M_1\rho_F + 2M_2}{\rho_M \rho_F}, \quad \text{and} \quad \rho \overset{def}{=} \min\{\rho_M, \rho_B, \rho_F\}.$$

With

$$c \geq \max\left\{ \frac{2\left(1 + \sqrt{1 + 8\Theta_1\Theta_2(2 - \rho_M + \rho_B\rho_M)}\right)}{2 - \rho_M + \rho_B\rho_M}, 16\Theta_1\Theta_2 \right\},$$

and $\nu \geq \max\left\{0, \frac{2-6\rho}{\rho}\right\}$, *set* $\gamma_k = \frac{k+\nu+4}{2cL}$ *and* $\tau_k = \frac{1}{cL\gamma_k}$. *After* T *iterations, Algorithm 1 produces a point* y_T *satisfying the following bound on its suboptimality:*

$$\mathbb{E}F(y_T) - F(x^*) \leq \frac{K_1(\nu+2)(\nu+4)}{(T+\nu+3)^2},$$

where

$$K_1 \overset{def}{=} F(y_0) - F(x^*) + \frac{2cL}{(\nu+2)(\nu+4)}\|z_0 - x^*\|^2.$$

A similar result gives an accelerated linear convergence rate when strong convexity is present.

Theorem 6 (**Acceleration With Strong Convexity**) *Suppose the stochastic gradient estimator* $\widetilde{\nabla}_{k+1}$ *satisfies the* MSEB$(M_1, M_2, \rho_M, \rho_B, \rho_F)$ *property and* g *is* μ*-strongly convex with* $\mu > 0$. *With the constants* Θ_1, Θ_2, c, *and* ν *set as in Theorem 5, set* $\gamma = \min\{\frac{1}{\sqrt{\mu c L}}, \frac{\rho}{2\mu}\}$ *and* $\tau = \mu\gamma$. *After* T *iterations, Algorithm 1 produces a point* z_T *satisfying the following bound:*

$$\mathbb{E}\|z_T - x^*\|^2 \leq K_2\left(1 + \min\left\{\sqrt{\frac{\mu}{Lc}}, \frac{\rho}{2}\right\}\right)^{-T}.$$

where

$$K_2 \overset{def}{=} \frac{2}{\mu}\left(F(y_0) - F(x^*)\right) + \|z_0 - x^*\|^2$$

Remark 2 *Although we prove accelerated convergence rates for many popular gradient estimators, the generality of Theorems 5 and 6 allows our results to extend easily to gradient estimators not considered in this work as well. These include, for example, the gradient estimators considered in [17].*

Remark 3 *With some manipulation, we see that these rates with* $c = \nu = \mathcal{O}(n)$ *are similar to the rates proved for Katyusha. In [2], the author shows that in the non-strongly convex case, Katyusha satisfies*

$$\mathbb{E}F(\widetilde{x}_S) - F(x^*) \leq \mathcal{O}\left(\frac{F(x_0) - F(x^*)}{S^2} + \frac{L\|x_0 - x^*\|^2}{PS^2}\right).$$

Recall that Katyusha follows the algorithmic framework of SVRG; S *denotes the epoch number,* $\widetilde{x}_S$ *the point where the full gradient was computed at the beginning of epoch* S, *and* $P = \mathcal{O}(n)$ *is the epoch length. In our notation,* $S = T/P = \mathcal{O}(T/n)$. *Theorem 5 with* $c = \nu = \mathcal{O}(n)$ *shows that Algorithm 1 achieves a similar convergence rate of*

$$\mathbb{E}F(y_T) - F(x^*) \leq \mathcal{O}\left(\frac{n^2}{T^2}\left(F(y_0) - F(x^*) + \frac{L}{n}\|z_0 - x^*\|^2\right)\right).$$

In the strongly convex case, an appropriately adapted version of Katyusha satisfies

$$\mathbb{E}F(\widetilde{x}_S) - F(x^*) \leq \begin{cases} \mathcal{O}\left(\left(1+\frac{\sqrt{\mu}}{\sqrt{LP}}\right)^{-SP}\right) & \frac{4}{3} \leq \frac{\sqrt{L}}{\sqrt{\mu n}} \\ \mathcal{O}\left(\left(\frac{3}{2}\right)^{-S}\right) & \frac{\sqrt{L}}{\sqrt{\mu n}} < \frac{4}{3}. \end{cases}$$

Similarly, with $c = \rho = \mathcal{O}(n)$, Theorem 6 shows that the iterates of Algorithm 1 satisfy

$$\frac{1}{2}\mathbb{E}\|z_T - x^*\|^2 \leq \mathcal{O}\left(\left(1+\min\left\{\sqrt{\frac{\mu}{Ln}}, \frac{1}{n}\right\}\right)^{-T}\right),$$

which again matches the rate of Katyusha. Of course, not all stochastic gradient estimators satisfy the bounds necessary to set $c = \nu = \rho = \mathcal{O}(n)$, so these optimal rates are conditional on being able to construct an "optimal estimator". SAGA, SVRG, SARAH, and SARGE all require c to be slightly larger than $\mathcal{O}(n)$.

The proofs of Theorems 5 and 6 use a linear coupling argument adapted from [6], but we use a different adaptation than the one in [2] used to prove convergence rates for Katyusha. To explain the differences between our approach and existing approaches, let us give a high-level description of linear coupling and the generalisation used in [2].

In [6], the authors suggest that gradient descent and mirror descent can be coupled to create an accelerated algorithm. We do not discuss gradient descent and mirror descent in detail (for this, see [6]), but the main idea of linear coupling can be understood from only two bounds arising from these algorithms. For the purpose of this argument, suppose $g \equiv 0$, so that $F \equiv f$. Gradient descent with step size η satisfies the following bound on the decrease of the objective (equation (2.1) in [6]):

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{\eta}\|\nabla f(x_{k+1})\|^2. \tag{6}$$

This bound shows that gradient descent is indeed a *descent method*; it is guaranteed to make progress at each iteration. The iterates of mirror descent using step size γ satisfy a bound on the sub-optimality of each iterate (equation (2.2) in [6]).

$$\langle \nabla f(x_k), x_k - x^* \rangle \leq \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 + \frac{\gamma^2}{2}\|\nabla f(x_k)\|^2. \tag{7}$$

While gradient descent is guaranteed to make progress proportional to $\|\nabla f(x_k)\|^2$ each iteration, mirror descent potentially introduces an "error" that is proportional to $\|\nabla f(x_k)\|^2$. Linear coupling takes advantage of this duality. Loosely speaking, by combining the sequence of iterates produced by gradient descent with the sequence produced by mirror descent, the guaranteed progress of gradient descent balances the potential error introduced by mirror descent, accelerating convergence.

This argument does not immediately hold for stochastic gradient methods. This is because in addition to the norm $\|\nabla f(x_k)\|^2$ arising in inequalities (6) and (7), we also get the MSE of our gradient estimator $\|\widetilde{\nabla} f(x_k) - \nabla f(x_k)\|^2$ as well as a "bias term". In the stochastic setting, analogues of inequalities (6) and (7) read

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle \\ &= f(x_k) - \frac{1}{\eta}\|\widetilde{\nabla} f(x_{k+1})\|^2 + \langle \nabla f(x_{k+1}) - \widetilde{\nabla} f(x_{k+1}), x_{k+1} - x_k \rangle \\ &\leq f(x_k) + \left(\frac{\epsilon}{2\eta^2} - \frac{1}{\eta}\right)\|\widetilde{\nabla} f(x_{k+1})\|^2 + \frac{1}{2\epsilon}\|\widetilde{\nabla} f(x_{k+1}) - \nabla f(x_{k+1})\|^2, \end{aligned}$$

where the last inequality is Young's, and

$$\gamma(f(x_k) - f(x^*)) \leq \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 + \gamma\left\langle \nabla f(x_k) - \widetilde{\nabla} f(x_k), x_k - x^* \right\rangle + \frac{\gamma^2}{2}\|\widetilde{\nabla} f(x_k)\|^2.$$

If the MSE or bias term is too large, the gradient step is no longer a descent step, and the progress does not balance the "error terms" in each of these inequalities, so we cannot expect linear coupling to offer any acceleration. This problem with the MSE and bias term exists for non-accelerated algorithms as well, and all analyses of stochastic gradient methods bound the effect of these terms, but in different ways. Katyusha and other accelerated algorithms in this family incorporate negative momentum to cancel part of the MSE. In contrast, analyses of non-accelerated algorithms do not try to cancel any of the variance, but show that the variance decreases fast enough so that it does not affect convergence rates.

3 Related Work

Besides Katyusha, there are many algorithms that use negative momentum for acceleration. In [30], the authors consider an accelerated version of SVRG that combines negative momentum with Nesterov's momentum to achieve the optimal $\sqrt{\kappa}$ dependence in the strongly convex case. This approach to acceleration is almost the same as Katyusha, but uses a traditional form of Nesterov's momentum instead of linear coupling. MiG [37] is another variant of these algorithms, corresponding to Katyusha with a certain parameter set to zero. VARAG is another approach to accelerated SVRG using negative momentum. VARAG achieves optimal convergence rates in the non-strongly convex and strongly convex settings under the framework of a single algorithm, and it converges linearly on problems that admit a global error bound, a quality that other algorithms have not yet been shown to possess [19].

The only direct acceleration of a SAGA-like algorithm is SSNM from [36]. Using the notation of (3), SSNM chooses a point from the set $\{\varphi_k^i\}_{i=1}^n$ uniformly at random, and uses this point as the "anchor point" for negative-momentum acceleration. Although SSNM admits fast convergence rates, there are a few undesirable qualities of this approach. SAGA has heavy storage requirements because it must store n gradients from previous iterations, and SSNM exacerbates this storage problem by storing n points from previous iterations as well. SSNM must also compute two stochastic gradients each iteration, so its per-iteration computational cost is similar to SVRG and Katyusha, and always higher than SAGA's.

Many algorithms for non-convex optimisation also use negative momentum for acceleration. KatyushaX [3] is a version of Katyusha adapted to optimise sum-of-non-convex objectives. To achieve its acceleration, KatyushaX uses classical momentum and a "retraction step", which is effectively an application of negative momentum (this relationship is acknowledged in [3] as well). Natasha [1] and Natasha2 [4] are accelerated algorithms for finding stationary points of non-convex objectives. Both algorithms employ a "retraction step" that is similar to negative momentum [1].

There are also many accelerated stochastic gradient algorithms that do not use negative momentum. In [24], the author applies Nesterov's momentum to SVRG without any sort of negative momentum, proving a linear convergence rate in the strongly convex regime. However, the proven convergence rate is suboptimal, as it implies even worse performance than SVRG when the batch size is small and worse performance than accelerated full-gradient methods when the batch size is close to n. Our results show that a particular application of Nesterov's momentum to SVRG does provide acceleration.

Point-SAGA [12] is another SAGA-like algorithm that achieves optimal convergence rates, but point-SAGA must compute the proximal operator corresponding to F rather than the proximal operator corresponding to g. This is not possible in general, even if the proximal operator corresponding to g is easy to compute, so point-SAGA applies to a different class of functions than the class we consider in this work.

There are also many algorithms that indirectly accelerate stochastic gradient methods. This class of algorithms include Catalyst [20], APPA [16], and the primal-dual methods in [35]. These algorithms call a variance-reduced stochastic gradient method as a subroutine, and provide acceleration using an inner-outer loop structure. These algorithms are often difficult to implement in practice due to the difficulty of solving their inner-loop subproblems, and they achieve a convergence rate that is only optimal up to a logarithmic factor.

4 Preliminaries

In this section, we present some basic definitions and results from optimisation and convex analysis. Much of our analysis involves *Bregman divergences*. The Bregman divergence associated with a function h is defined

as

$$D_h^\xi(y,x) \stackrel{\text{def}}{=} h(y) - h(x) + \langle \xi, x - y \rangle,$$

where $\xi \in \partial h(x)$ and ∂ is the subdifferential operator. If h is differentiable, we drop the superscript ξ as the subgradient is unique. The function h is convex if and only if $D_h^\xi(y,x) \geq 0$ for all x and y. We say h is *μ-strongly convex* with $\mu \geq 0$ if and only if

$$\frac{\mu}{2}\|x - y\|^2 \leq D_h^\xi(y,x).$$

Bregman divergences also arise in the following fundamental inequality.

Lemma 7 ([22], Thm. 2.1.5) *Suppose f is convex with an L-Lipschitz continuous gradient. We have for all $x, y \in \mathbb{R}^m$,*

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2LD_f(y,x).$$

Lemma 7 is equivalent to the following result, which is more specific to our analysis due to the finite-sum structure of the smooth term in (1).

Lemma 8 *Let $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, where each f_i is convex with an L-Lipschitz continuous gradient. Then for every $x, y \in \mathbb{R}^m$*

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2LD_f(y,x).$$

Proof. This follows from applying Lemma 7 to each component f_i. □

The *proximal operator* is defined as

$$\text{prox}_g(y) = \arg\min_{x \in \mathbb{R}^m} \left\{ \frac{1}{2}\|x - y\|^2 + g(x) \right\}.$$

The proximal operator is also defined implicitly as $y - \text{prox}_g(y) \in \partial g(\text{prox}_g\,(y))$. From this definition of the proximal operator, the following standard inequality is clear.

Lemma 9 *Suppose g is μ-strongly convex with $\mu \geq 0$, and suppose $z = \text{prox}_{\eta g}\,(x - \eta d)$ for some $x, d \in \mathbb{R}^m$ and constant η. Then, for any $y \in \mathbb{R}^m$,*

$$\eta\langle d, z - y\rangle \leq \frac{1}{2}\|x - y\|^2 - \frac{1 + \mu\eta}{2}\|z - y\|^2 - \frac{1}{2}\|z - x\|^2 - \eta g(z) + \eta g(y).$$

Proof. By the strong convexity of g,

$$g(z) - g(y) \leq \langle \xi, z - y\rangle - \frac{\mu}{2}\|z - y\|^2 \quad \forall \xi \in \partial g(z)$$

From the implicit definition of the proximal operator, we know that $\frac{1}{\eta}(z - x) + d \in \partial g(z)$. Therefore,

$$\begin{aligned} g(z) - g(y) &\leq \langle \xi, z - y\rangle - \frac{\mu}{2}\|z - y\|^2 \\ &= \frac{1}{\eta}\langle z - x + \eta d, z - y\rangle - \frac{\mu}{2}\|z - y\|^2 \\ &= \langle d, z - y\rangle + \frac{1}{2\eta}\|x - y\|^2 - \frac{1 + \mu\eta}{2\eta}\|z - y\|^2 - \frac{1}{2\eta}\|z - x\|^2. \end{aligned}$$

Multiplying by η and rearranging yields the assertion. □

5 The Acceleration Framework

To apply the linear coupling framework, we must couple stochastic analogues of (6) and (7) to construct a lower bound on the one-iteration progress of Algorithm 1.

Lemma 10 (**One-Iteration Progress**) *The following bound describes the progress made by one iteration of Algorithm 1.*

$$\begin{aligned}
0 \leq & \frac{\gamma_k(1-\tau_k)}{\tau_k}F(y_k) - \frac{\gamma_k}{\tau_k}F(y_{k+1}) + \gamma_k F(x^*) + \gamma_k^2\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 \\
& + \frac{\gamma_k}{\tau_k}\left(\frac{L}{2} - \frac{1}{4\tau_k\gamma_k}\right)\|x_{k+1} - y_{k+1}\|^2 + \frac{1}{2}\|z_k - x^*\|^2 \\
& - \frac{1+\mu\gamma_k}{2}\|z_{k+1} - x^*\|^2 + \gamma_k\left\langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_k - x^*\right\rangle \\
& - \frac{\gamma_k(1-\tau_k)}{\tau_k}D_f(y_k, x_{k+1}).
\end{aligned}$$

Proof. We use a linear coupling argument. The extrapolated iterate x_{k+1} can be viewed as a convex combination of an iterate produced from mirror descent (namely, z_k) and one from gradient descent (y_k). This allows us to provide two bounds on the term $f(x_{k+1}) - f(x^*)$: one is a regret bound inspired by the classical analysis of mirror descent, and the other is inspired by the traditional descent guarantee of gradient descent.

$$\begin{aligned}
& \gamma_k(f(x_{k+1}) - f(x^*)) \\
& \overset{①}{\leq} \gamma_k\langle \nabla f(x_{k+1}), x_{k+1} - x^*\rangle \\
& = \gamma_k\langle \nabla f(x_{k+1}), x_{k+1} - z_k\rangle + \langle \nabla f(x_{k+1}), z_k - x^*\rangle \\
& \overset{②}{=} \frac{\gamma_k(1-\tau_k)}{\tau_k}\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle + \gamma_k\langle \nabla f(x_{k+1}), z_k - x^*\rangle \\
& = \frac{\gamma_k(1-\tau_k)}{\tau_k}(f(y_k) - f(x_{k+1})) + \gamma_k\left\langle \widetilde{\nabla}_{k+1}, z_k - x^*\right\rangle \\
& \quad - \frac{\gamma_k(1-\tau_k)}{\tau_k}D_f(y_k, x_{k+1}) + \gamma_k\left\langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_k - x^*\right\rangle \\
& = \frac{\gamma_k(1-\tau_k)}{\tau_k}(f(y_k) - f(x_{k+1})) + \gamma_k\left\langle \widetilde{\nabla}_{k+1}, z_k - z_{k+1}\right\rangle \\
& \quad + \gamma_k\left\langle \widetilde{\nabla}_{k+1}, z_{k+1} - x^*\right\rangle - \frac{\gamma_k(1-\tau_k)}{\tau_k}D_f(y_k, x_{k+1}) \\
& \quad + \gamma_k\left\langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_k - x^*\right\rangle \\
& \overset{③}{=} \frac{\gamma_k(1-\tau_k)}{\tau_k}(f(y_k) - f(x_{k+1})) + \frac{\gamma_k}{\tau_k}\left\langle \widetilde{\nabla}_{k+1}, x_{k+1} - y_{k+1}\right\rangle \\
& \quad + \gamma_k\left\langle \widetilde{\nabla}_{k+1}, z_{k+1} - x^*\right\rangle - \frac{\gamma_k(1-\tau_k)}{\tau_k}D_f(y_k, x_{k+1}) \\
& \quad + \gamma_k\left\langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_k - x^*\right\rangle
\end{aligned} \tag{8}$$

Inequality ① uses the convexity of f, ② follows from the fact that $x_{k+1} = \tau_k z_k + (1-\tau_k)y_k$, and ③ uses $x_{k+1} - y_{k+1} = \tau_k(z_k - z_{k+1})$. We proceed to bound the inner product $\langle \widetilde{\nabla}_{k+1}, z_{k+1} - x^*\rangle$ involving the sequence z_{k+1} using a regret bound from mirror descent, and we bound the term $\langle \widetilde{\nabla}_{k+1}, x_{k+1} - y_{k+1}\rangle$ using an argument similar to the descent guarantee of gradient descent.

By Lemma 9 with $z = z_{k+1}$, $x = z_k$, $y = x^*$, $d = \widetilde{\nabla}_{k+1}$, and $\eta = \gamma_k$,

$$\begin{aligned}
& \gamma_k\left\langle \widetilde{\nabla}_{k+1}, z_{k+1} - x^*\right\rangle \\
& \leq \frac{1}{2}\|z_k - x^*\|^2 - \frac{1+\mu\gamma_k}{2}\|z_{k+1} - x^*\|^2 - \frac{1}{2}\|z_{k+1} - z_k\|^2
\end{aligned}$$

$$
\begin{aligned}
&\quad - \gamma_k g(z_{k+1}) + \gamma_k g(x^*) \\
&= \frac{1}{2}\|z_k - x^*\|^2 - \frac{1+\mu\gamma_k}{2}\|z_{k+1} - x^*\|^2 - \frac{1}{2\tau_k^2}\|x_{k+1} - y_{k+1}\|^2 \\
&\quad - \gamma_k g(z_{k+1}) + \gamma_k g(x^*). \qquad (9)
\end{aligned}
$$

For the other term,

$$
\begin{aligned}
&\frac{\gamma_k}{\tau_k}\langle \widetilde{\nabla}_{k+1}, x_{k+1} - y_{k+1}\rangle \\
=&\frac{\gamma_k}{\tau_k}\langle \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle + \frac{\gamma_k}{\tau_k}\langle \widetilde{\nabla}_{k+1} - \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle \\
\overset{①}{\leq}&\frac{\gamma_k}{\tau_k}\left(f(x_{k+1}) - f(y_{k+1})\right) + \frac{\gamma_k}{\tau_k}\langle \widetilde{\nabla}_{k+1} - \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle \\
&+ \frac{L\gamma_k}{2\tau_k}\|x_{k+1} - y_{k+1}\|^2 \\
\overset{②}{\leq}&\frac{\gamma_k}{\tau_k}\left(f(x_{k+1}) - f(y_{k+1})\right) + \gamma_k^2\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 \\
&+ \left(\frac{L\gamma_k}{2\tau_k} + \frac{1}{4\tau_k^2}\right)\|x_{k+1} - y_{k+1}\|^2 \\
=&\frac{\gamma_k}{\tau_k}\left(f(x_{k+1}) - F(y_{k+1})\right) + \gamma_k^2\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 \\
&+ \left(\frac{L\gamma_k}{2\tau_k} + \frac{1}{4\tau_k^2}\right)\|x_{k+1} - y_{k+1}\|^2 + \frac{\gamma_k}{\tau_k}g(y_{k+1}) \\
\overset{③}{\leq}&\frac{\gamma_k}{\tau_k}\left(f(x_{k+1}) - F(y_{k+1})\right) + \gamma_k^2\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 \\
&+ \left(\frac{L\gamma_k}{2\tau_k} + \frac{1}{4\tau_k^2}\right)\|x_{k+1} - y_{k+1}\|^2 + \gamma_k g(z_{k+1}) + \frac{\gamma_k(1-\tau_k)}{\tau_k}g(y_k). \qquad (10)
\end{aligned}
$$

Inequality ① follows from the Lipschitz continuity of ∇f_i, ② is Young's inequality, and ③ uses the convexity of g and the update rule $y_{k+1} = \tau_k z_{k+1} + (1-\tau_k)y_k$. Combining inequalities (9) and (10) with (8) and rearranging yields the assertion. □

Lemma 10 completes the linear coupling part of our argument. If not for the MSE and bias terms, we could telescope this inequality as in [6] and prove an accelerated convergence rate. As with all analyses of stochastic gradient methods, we need a useful bound on these qualities of the estimator.

Existing analyses of unbiased stochastic gradient methods bound the variance term by a pair of terms that telescope over several iterations, showing that the variance tends to zero with the number of iterations. It is difficult to generalise these arguments to accelerated stochastic methods because one must prove that the variance decreases at an accelerated rate that is inconsistent with existing variance bounds. In the analysis of Katyusha, negative momentum cancels part of the variance term, leaving telescoping terms that decrease at an accelerated rate. Without negative momentum, we must handle the variance term differently.

In the inequality of Lemma 10, we have two non-positive terms:

$$
-\frac{1}{\tau_k^2}\|x_{k+1} - y_{k+1}\|^2 \quad \text{and} \quad -\frac{\gamma_k(1-\tau_k)}{\tau_k}D_f(y_k, x_{k+1}).
$$

This makes our strategy clear: we must bound the MSE and bias terms by terms of the form $\|x_{k+1} - y_{k+1}\|^2$ and $D_f(y_k, x_{k+1})$. The following two lemmas use the MSEB property to establish bounds of this form.

Lemma 11 (**Bias Term Bound**) *Suppose the stochastic gradient estimator $\widetilde{\nabla}_{k+1}$ satisfies the* MSEB(M_1, M_2, ρ_M, ρ_B, ρ_F) *property, let $\rho = \min\{\rho_M, \rho_B, \rho_F\}$, and let $\{\sigma_k\}$ and $\{s_k\}$ be any non-negative sequences satisfying $\sigma_k s_k^2 (1-\rho) \leq \sigma_{k-1}s_{k-1}^2\left(1 - \frac{\rho}{2}\right)$ and $\sigma_k(1-\rho) \leq \sigma_{k-1}\left(1 - \frac{\rho}{2}\right)$. The bias term can be bounded as*

$$
\sum_{k=0}^{T-1} \sigma_k s_k \mathbb{E}\left\langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_k - x^* \right\rangle
$$

$$\leq (1-\rho_B)\sum_{k=0}^{T-1}\sigma_k\mathbb{E}\left[\frac{8s_k^2}{\rho_B^2\rho_M}\left\|\nabla f(x_{k+1})-\widetilde{\nabla}_{k+1}\right\|^2+\frac{\rho_M}{8\tau_k^2}\|x_{k+1}-y_{k+1}\|^2\right].$$

Proof. Because z_k is independent of the first $k-1$ iterates, we can use the MSEB property to say

$$\begin{aligned}
&\sigma_k s_k\mathbb{E}\left\langle \nabla f(x_{k+1})-\widetilde{\nabla}_{k+1}, z_k-x^*\right\rangle\\
=&\sigma_k s_k\mathbb{E}\left\langle \nabla f(x_{k+1})-\mathbb{E}_k\widetilde{\nabla}_{k+1}, z_k-x^*\right\rangle\\
\overset{①}{=}&\sigma_k s_k(1-\rho_B)\mathbb{E}\left\langle \nabla f(x_k)-\widetilde{\nabla}_k, z_k-x^*\right\rangle\\
\overset{②}{=}&\sigma_k(1-\rho_B)\mathbb{E}\Big[s_k\left\langle \nabla f(x_k)-\widetilde{\nabla}_k, z_k-z_{k-1}\right\rangle+s_k\left\langle \nabla f(x_k)-\mathbb{E}_{k-1}\widetilde{\nabla}_k, z_{k-1}-x^*\right\rangle\Big]\\
\overset{③}{\leq}&\sigma_k(1-\rho_B)\mathbb{E}\left[\frac{4s_k^2}{\rho_M\rho_B}\left\|\nabla f(x_k)-\widetilde{\nabla}_k\right\|^2+\frac{\rho_M\rho_B}{16}\|z_k-z_{k-1}\|^2+s_k\left\langle \nabla f(x_k)-\mathbb{E}_{k-1}\widetilde{\nabla}_k, z_{k-1}-x^*\right\rangle\right].
\end{aligned}$$

Equality ① is due to the MSEB property. We are able to pass the conditional expectation into the second inner product in ② because z_{k-1} is independent of $\widetilde{\nabla}_k$ conditioned on the first $k-2$ iterates, and inequality ③ is Young's. We can repeat this process once more, applying the MSEB property to obtain

$$\begin{aligned}
&\sigma_k(1-\rho_B)\mathbb{E}\Bigg[\frac{4s_k^2}{\rho_M\rho_B}\left\|\nabla f(x_k)-\widetilde{\nabla}_k\right\|^2+\frac{\rho_M\rho_B}{16}\|z_k-z_{k-1}\|^2\\
&\quad+s_k(1-\rho_B)\left\langle \nabla f(x_{k-1})-\widetilde{\nabla}_{k-1}, z_{k-1}-x^*\right\rangle\Bigg]\\
\leq&(1-\rho_B)\mathbb{E}\Bigg[\frac{4\sigma_k s_k^2}{\rho_M\rho_B}\left\|\nabla f(x_k)-\widetilde{\nabla}_k\right\|^2+\sigma_k s_k^2(1-\rho_B)\left\|\nabla f(x_{k-1})-\widetilde{\nabla}_{k-1}\right\|^2\\
&\quad+\frac{\rho_M\rho_B}{16}\left(\sigma_k\|z_k-z_{k-1}\|^2+\sigma_k(1-\rho_B)\|z_{k-1}-z_{k-2}\|^2\right)\\
&\quad+\sigma_k s_k(1-\rho_B)\left\langle \nabla f(x_{k-1})-\widetilde{\nabla}_{k-1}, z_{k-1}-x^*\right\rangle\Bigg]\\
\overset{④}{\leq}&(1-\rho_B)\mathbb{E}\Bigg[\frac{4\sigma_k s_k^2}{\rho_M\rho_B}\left\|\nabla f(x_k)-\widetilde{\nabla}_k\right\|^2\\
&\quad+\sigma_{k-1}s_{k-1}^2\left(1-\frac{\rho_B}{2}\right)\left\|\nabla f(x_{k-1})-\widetilde{\nabla}_{k-1}\right\|^2+\frac{\rho_M\rho_B}{16}\Big(\sigma_k\|z_k-z_{k-1}\|^2\\
&\quad+\sigma_{k-1}\left(1-\frac{\rho_B}{2}\right)\|z_{k-1}-z_{k-2}\|^2\Big)\\
&\quad+\sigma_k s_k(1-\rho_B)\left\langle \nabla f(x_{k-1})-\widetilde{\nabla}_{k-1}, z_{k-1}-x^*\right\rangle\Bigg].
\end{aligned}$$

Inequality ④ uses our hypothesis on the decrease of s_k. This is a recursive inequality, and expanding the recursion yields

$$\begin{aligned}
&\sigma_k s_k\mathbb{E}\left\langle \nabla f(x_{k+1})-\mathbb{E}_k\widetilde{\nabla}_{k+1}, z_k-x^*\right\rangle\\
\leq&(1-\rho_B)\sum_{\ell=1}^{k}\sigma_\ell\mathbb{E}\left[\frac{4s_\ell^2(1-\frac{\rho_B}{2})^{k-\ell}}{\rho_M\rho_B}\left\|\nabla f(x_\ell)-\widetilde{\nabla}_\ell\right\|^2+\frac{\rho_M\rho_B(1-\frac{\rho_B}{2})^{k-\ell}}{16}\|z_\ell-z_{\ell-1}\|^2\right].
\end{aligned}$$

Taking the sum over the iterations $k=0$ to $k=T-1$, we apply an estimate to simplify this bound.

$$\sum_{k=0}^{T-1}\sigma_k s_k\mathbb{E}\left\langle \nabla f(x_{k+1})-\mathbb{E}_k\widetilde{\nabla}_{k+1}, z_k-x^*\right\rangle$$

$$
\begin{aligned}
&\leq (1-\rho_B)\sum_{k=0}^{T-1}\sum_{\ell=1}^{k}\sigma_\ell\mathbb{E}\Big[\frac{4s_\ell^2(1-\frac{\rho_B}{2})^{k-\ell}}{\rho_M\rho_B}\left\|\nabla f(x_\ell)-\widetilde{\nabla}_\ell\right\|^2+\frac{\rho_M\rho_B(1-\frac{\rho_B}{2})^{k-\ell}}{16}\|z_\ell-z_{\ell-1}\|^2\Big]\\
&\leq (1-\rho_B)\sum_{k=0}^{T-1}\sigma_k\left(\sum_{\ell=1}^{\infty}\left(1-\frac{\rho_B}{2}\right)^\ell\right)\mathbb{E}\left[\frac{4s_k^2}{\rho_M\rho_B}\left\|\nabla f(x_{k+1})-\widetilde{\nabla}_{k+1}\right\|^2+\frac{\rho_M\rho_B}{16}\|z_{k+1}-z_k\|^2\right] \qquad (11)\\
&= (1-\rho_B)\sum_{k=0}^{T-1}\sigma_k\mathbb{E}\left[\frac{8s_k^2}{\rho_B^2\rho_M}\left\|\nabla f(x_{k+1})-\widetilde{\nabla}_{k+1}\right\|^2+\frac{\rho_M}{8}\|z_{k+1}-z_k\|^2\right]\\
&\overset{①}{=} (1-\rho_B)\sum_{k=0}^{T-1}\sigma_k\mathbb{E}\left[\frac{8s_k^2}{\rho_B^2\rho_M}\left\|\nabla f(x_{k+1})-\widetilde{\nabla}_{k+1}\right\|^2+\frac{\rho_M}{8\tau_k^2}\|x_{k+1}-y_{k+1}\|^2\right].
\end{aligned}
$$

Equality ① is the identity $y_{k+1}-x_{k+1}=\tau_k(z_{k+1}-z_k)$. □

This bound on the bias term includes the MSE, so to complete our bound on the bias term, we must combine Lemma 11 with the following lemma.

Lemma 12 (**MSE Bound**) *Suppose the stochastic gradient estimator $\widetilde{\nabla}_{k+1}$ satisfies the* MSEB$(M_1, M_2, \rho_M, \rho_B, \rho_F)$ *property, let $\rho=\min\{\rho_M,\rho_B,\rho_F\}$, and let $\{s_k\}$ be any non-negative sequence satisfying $s_k^2(1-\rho)\leq s_{k-1}^2\left(1-\frac{\rho}{2}\right)$. For convenience, define $\Theta_2=\frac{M_1\rho_F+2M_2}{\rho_M\rho_F}$. The MSE of the gradient estimator is bounded as*

$$
\sum_{k=0}^{T-1}s_k^2\mathbb{E}\|\nabla f(x_{k+1})-\widetilde{\nabla}_{k+1}\|^2\leq\sum_{k=0}^{T-1}4\Theta_2 L s_k^2\mathbb{E}\left[2D_f(y_k,x_{k+1})+L\|x_{k+1}-y_{k+1}\|^2\right]
$$

Proof. First, we derive a bound on the sequence $\mathcal{F}_k$ arising in the MSEB property. Taking the sum from $k=0$ to $k=T-1$,

$$
\begin{aligned}
\sum_{k=0}^{T-1}s_k^2\mathcal{F}_k&\leq\sum_{k=0}^{T-1}\sum_{\ell=0}^{k}\frac{M_2s_k^2(1-\rho_F)^{k-\ell}}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{\ell+1})-\nabla f_i(x_\ell)\|^2\\
&\overset{①}{\leq}\sum_{k=0}^{T-1}\sum_{\ell=0}^{k}\frac{M_2s_\ell^2(1-\frac{\rho_F}{2})^{k-\ell}}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{\ell+1})-\nabla f_i(x_\ell)\|^2\\
&\overset{②}{\leq}\sum_{k=0}^{T-1}\frac{2M_2s_k^2}{n\rho_F}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2.
\end{aligned}
$$

Inequality ① uses the fact that $s_k^2(1-\rho_F)\leq s_{k-1}^2\left(1-\frac{\rho_F}{2}\right)$, and ② uses the same estimate as in (11). With this bound on $\mathcal{F}_k$, we proceed to bound $\mathcal{M}_k$ in a similar fashion.

$$
\begin{aligned}
&\sum_{k=0}^{T-1}s_k^2\mathbb{E}\|\nabla f(x_{k+1})-\widetilde{\nabla}_{k+1}\|^2\\
&\leq\sum_{k=0}^{T-1}\frac{M_1s_k^2}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2+s_k^2\mathcal{F}_k+s_k^2(1-\rho_M)\mathcal{M}_{k-1}\\
&\leq\sum_{k=0}^{T-1}\frac{(M_1\rho_F+2M_2)s_k^2}{n\rho_F}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2+s_k^2(1-\rho_M)\mathcal{M}_{k-1}\\
&\leq\sum_{k=0}^{T-1}\sum_{\ell=1}^{k}\frac{\Theta_2s_k^2(1-\rho_M)^{k-\ell}\rho_M}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{\ell+1})-\nabla f_i(x_\ell)\|^2\\
&\overset{①}{\leq}\sum_{k=0}^{T-1}\sum_{\ell=1}^{k}\frac{\Theta_2s_\ell^2(1-\frac{\rho_M}{2})^{k-\ell}\rho_M}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{\ell+1})-\nabla f_i(x_\ell)\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\overset{②}{\leq} \sum_{k=0}^{T-1} \frac{2\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E}\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|^2 \\
&\overset{③}{\leq} \sum_{k=0}^{T-1} \frac{4\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E}\|\nabla f_i(x_{k+1}) - \nabla f_i(y_k)\|^2 + \frac{4\Theta_2 s_k^2}{n} \sum_{i=1}^{n} \mathbb{E}\|\nabla f_i(y_k) - \nabla f_i(x_k)\|^2 \\
&\overset{④}{\leq} \sum_{k=0}^{T-1} \left(8\Theta_2 L s_k^2 \mathbb{E} D_f(y_k, x_{k+1}) + 4\Theta_2 L^2 s_k^2 \mathbb{E}\|x_k - y_k\|^2\right).
\end{aligned}
$$

Inequality ① uses $s_k^2(1-\rho_M) \leq s_{k-1}^2\left(1-\frac{\rho_M}{2}\right)$, ② uses the same estimate we applied above, ③ uses the inequality $\|a-c\|^2 \leq 2\|a-b\|^2 + 2\|b-c\|^2$, and ④ uses Lemma 7 and the Lipschitz continuity of ∇f_i. □

Lemmas 11 and 12 show that it is possible to cancel the bias term and the MSE using the non-negative terms appearing in the inequality of Lemma 10. Without these terms, we can telescope this inequality over several iterations and prove accelerated convergence rates. We are now prepared to prove Theorems 5 and 6.

Proof of Theorem 5. We set $\mu = 0$ in the inequality of Lemma 10, apply the full expectation operator, and sum the result over the iterations $k = 0$ to $k = T-1$.

$$
\begin{aligned}
0 \leq& \frac{1}{2}\|z_0 - x^*\|^2 - \frac{1}{2}\mathbb{E}\|z_T - x^*\|^2 + \sum_{k=0}^{T-1} \mathbb{E}\Big[\frac{\gamma_k(1-\tau_k)}{\tau_k} F(y_k) - \frac{\gamma_k}{\tau_k} F(y_{k+1}) \\
&+ \gamma_k F(x^*) + \frac{\gamma_k}{\tau_k}\left(\frac{L}{2} - \frac{1}{4\tau_k\gamma_k}\right)\|x_{k+1} - y_{k+1}\|^2 - \frac{\gamma_k(1-\tau_k)}{\tau_k} D(y_k, x_{k+1}) \\
&+ \gamma_k \left\langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_k - x^* \right\rangle + \gamma_k^2 \|\nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}\|^2\Big].
\end{aligned}
$$

We bound the terms in the final line, beginning with the bias term. Our choice for γ_k satisfies $\gamma_k^2(1-\rho) \leq \gamma_{k-1}^2\left(1-\frac{\rho}{2}\right)$, so with $s_k = \gamma_k$ and $\sigma_k = 1$, we apply Lemma 11. This gives

$$
\begin{aligned}
0 \leq& \frac{1}{2}\|z_0 - x^*\|^2 + \sum_{k=0}^{T-1} \mathbb{E}\left[\frac{\gamma_k(1-\tau_k)}{\tau_k} F(y_k) - \frac{\gamma_k}{\tau_k} F(y_{k+1}) + \gamma_k F(x^*)\right. \\
&+ \left(\frac{\gamma_k}{\tau_k}\left(\frac{L}{2} - \frac{1}{4\tau_k\gamma_k}\right) + \frac{(1-\rho_B)\rho_M}{8\tau_k^2}\right)\|x_{k+1} - y_{k+1}\|^2 \\
&\left.- \frac{\gamma_k(1-\tau_k)}{\tau_k} D(y_k, x_{k+1}) + \gamma_k^2\Theta_1\|\nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}\|^2\right],
\end{aligned}
$$

where we have dropped the term $-1/2\mathbb{E}\|z_T - x^*\|^2$ because it is non-positive. Applying Lemma 12 to bound the MSE, we have

$$
\begin{aligned}
0 \leq& \frac{1}{2}\|z_0 - x^*\|^2 + \sum_{k=0}^{T-1} \mathbb{E}\Big[\frac{\gamma_k(1-\tau_k)}{\tau_k} F(y_k) - \frac{\gamma_k}{\tau_k} F(y_{k+1}) \\
&+ \gamma_k F(x^*) + \left(8\gamma_k^2 L\Theta_1\Theta_2 - \frac{\gamma_k(1-\tau_k)}{\tau_k}\right) D(y_k, x_{k+1}) \\
&+ \left(\frac{\rho_M(1-\rho_B)}{8\tau_k^2} + 4\gamma_k^2 L^2\Theta_1\Theta_2 + \frac{\gamma_k}{\tau_k}\left(\frac{L}{2} - \frac{1}{4\tau_k\gamma_k}\right)\right)\|x_{k+1} - y_{k+1}\|^2\Big].
\end{aligned} \tag{12}
$$

With the parameters set as in the theorem statement, it is clear that the final two lines of (12) are non-positive (see Appendix A for a proof). This allows us to drop these lines from the inequality, leaving

$$
0 \leq \frac{1}{2}\|z_0 - x^*\|^2 + \sum_{k=0}^{T-1} \mathbb{E}\left[\frac{\gamma_k(1-\tau_k)}{\tau_k} F(y_k) - \frac{\gamma_k}{\tau_k} F(y_{k+1}) + \gamma_k F(x^*)\right].
$$

Rewriting τ_k in terms of γ_k shows that this is equivalent to

$$0 \leq \frac{1}{2}\|z_0 - x^*\|^2 - \frac{1}{2}\mathbb{E}\|z_T - x^*\|^2 + \sum_{k=0}^{T-1} \mathbb{E}\left[(cL\gamma_k^2 - \gamma_k)F(y_k) - cL\gamma_k^2 F(y_{k+1}) + \gamma_k F(x^*)\right].$$

Our choice for γ_k satisfies $cL\gamma_k^2 - \gamma_k = cL\gamma_{k-1}^2 - \frac{1}{4cL}$, allowing the $F(y_k)$ terms to telescope. Hence, our inequality is equivalent to

$$0 \leq -cL\gamma_{T-1}^2\mathbb{E}[F(y_T) - F(x^*)] - \frac{1}{4cL}\sum_{k=1}^{T-1}\mathbb{E}\left[F(y_k) - F(x^*)\right] + (cL\gamma_0^2 - \gamma_0)(F(y_0) - F(x^*)) + \frac{1}{2}\|z_0 - x^*\|^2.$$

Using the facts that $cL\gamma_{T-1}^2 = \frac{(T+\nu+3)^2}{4cL}$, $cL\gamma_0^2 - \gamma_0 = \frac{(\nu+2)(\nu+4)}{4cL}$, and $F(y_k) \leq F(x^*)$, we have

$$\frac{(T+\nu+3)^2}{4cL}\mathbb{E}[F(y_T) - F(x^*)] \leq \frac{(\nu+2)(\nu+4)}{4cL}(F(y_0) - F(x^*)) + \frac{1}{2}\|z_0 - x^*\|^2.$$

Thie proves the assertion.

□

A similar argument proves an accelerated linear convergence rate when strong convexity is present.

Proof of Theorem 6. We recall the inequality of Lemma 10.

$$\begin{aligned}
&\frac{\gamma}{\tau}\left(F(y_{k+1}) - F(x^*)\right) + \frac{(1+\mu\gamma)}{2}\|z_{k+1} - x^*\|^2 \\
\leq &\frac{\gamma(1-\tau)}{\tau}\left(F(y_k) - F(x^*)\right) + \frac{1}{2}\|z_k - x^*\|^2 + \gamma^2\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 \\
&+ \frac{\gamma}{\tau}\left(\frac{L}{2} - \frac{1}{4\tau\gamma}\right)\|x_{k+1} - y_{k+1}\|^2 - \frac{\gamma(1-\tau)}{\tau}D_f(y_k, x_{k+1}) \\
&+ \gamma\left\langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_k - x^*\right\rangle.
\end{aligned}$$

By our choice of γ and τ, we have

$$\frac{\gamma}{\tau}\left(\frac{\gamma(1-\tau)}{\tau}\right)^{-1} = \frac{1}{1-\tau} \geq 1 + \tau = 1 + \mu\gamma.$$

Therefore, we can extract a factor of $(1+\mu\gamma)$ from the left.

$$\begin{aligned}
&(1+\mu\gamma)\left(\frac{\gamma(1-\tau)}{\tau}\left(F(y_{k+1}) - F(x^*)\right) + \frac{1}{2}\|z_{k+1} - x^*\|^2\right) \\
\leq &\frac{\gamma(1-\tau)}{\tau}\left(F(y_k) - F(x^*)\right) + \frac{1}{2}\|z_k - x^*\|^2 + \gamma^2\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1})\|^2 \\
&+ \frac{\gamma}{\tau}\left(\frac{L}{2} - \frac{1}{4\tau\gamma}\right)\|x_{k+1} - y_{k+1}\|^2 - \frac{\gamma(1-\tau)}{\tau}D_f(y_k, x_{k+1}) \\
&+ \gamma\left\langle \nabla f(x_{k+1}) - \widetilde{\nabla}_{k+1}, z_k - x^*\right\rangle.
\end{aligned}$$

Multiplying this inequality by $(1+\mu\gamma)^k$, summing over iterations $k = 0$ to $k = T-1$, and applying the full expectation operator, we obtain the bound

$$(1+\mu\gamma)^T\mathbb{E}\left[\frac{\gamma(1-\tau)}{\tau}\left(F(y_T) - F(x^*)\right) + \frac{1}{2}\|z_T - x^*\|^2\right]$$

$$\leq \frac{\gamma(1-\tau)}{\tau}\left(F(y_0)-F(x^*)\right)+\frac{1}{2}\|z_0-x^*\|^2+\sum_{k=0}^{T-1}(1+\mu\gamma)^k\mathbb{E}\Big[\gamma^2\|\widetilde{\nabla}_{k+1}-\nabla f(x_{k+1})\|^2 \tag{13}$$
$$+\frac{\gamma}{\tau}\left(\frac{L}{2}-\frac{1}{4\tau\gamma}\right)\|x_{k+1}-y_{k+1}\|^2-\frac{\gamma(1-\tau)}{\tau}D_f(y_k,x_{k+1})+\gamma\left\langle\nabla f(x_{k+1})-\widetilde{\nabla}_{k+1},z_k-x^*\right\rangle\Big].$$

As in the proof of Theorem 5, we bound the bias term and the MSE using Lemmas 11 and 12, respectively. To apply Lemma 11, we let $\sigma_k=(1+\mu\gamma)^k$ and $s_k=\gamma$. These choices are appropriate because $(1+\mu\gamma)^k(1-\rho)\leq(1+\mu\gamma)^{k-1}(1-\frac{\rho}{2})$ due to the fact that $\mu\gamma\leq\rho/2$.

Combining these bounds with (13), we have

$$\begin{aligned}
&(1+\mu\gamma)^T\mathbb{E}\left[\frac{\gamma(1-\tau)}{\tau}\left(F(y_T)-F(x^*)\right)+\frac{1}{2}\|z_T-x^*\|^2\right]\\
\leq&\frac{\gamma(1-\tau)}{\tau}\left(F(y_0)-F(x^*)\right)+\frac{1}{2}\|z_0-x^*\|^2\\
&+\sum_{k=0}^{T-1}(1+\mu\gamma)^k\mathbb{E}\Bigg[\left(8\gamma^2L\Theta_1\Theta_2-\frac{\gamma(1-\tau)}{\tau}\right)D(y_k,x_{k+1})\\
&+\left(\frac{\rho_M(1-\rho_B)}{8\tau^2}+4\gamma^2L^2\Theta_1\Theta_2+\frac{\gamma}{\tau}\left(\frac{L}{2}-\frac{1}{4\tau\gamma}\right)\right)\|x_{k+1}-y_{k+1}\|^2\Big].
\end{aligned}$$

The parameter settings in the theorem statement ensure the final two lines are non-positive (see Appendix A for details). This gives us

$$\begin{aligned}
\frac{1}{2}\mathbb{E}\|z_T-x^*\|^2&\leq(1+\mu\gamma)^{-T}\left(\frac{\gamma(1-\tau)}{\tau}\left(F(y_0)-F(x^*)\right)+\frac{1}{2}\|z_0-x^*\|^2\right)\\
&\leq\left(1+\min\left\{\sqrt{\frac{\mu}{Lc}},\frac{\rho}{2}\right\}\right)^{-T}\left(\frac{1}{\mu}\left(F(y_0)-F(x^*)\right)+\frac{1}{2}\|z_0-x^*\|^2\right),
\end{aligned}$$

which is the desired result. □

6 Convergence Rates for Specific Estimators

In light of Theorems 5 and 6, we must only establish suitable bounds on the MSE and bias terms of a gradient estimator to prove accelerated convergence rates for Algorithm 1. We consider four variance-reduced gradient estimators: SAGA, SVRG, SARAH, and SARGE, beginning with the unbiased estimators. We defer proofs to the Appendix. To preserve the generality of our framework, we have not optimised the constants appearing in the presented convergence rates.

Theorem 13 (SAGA Convergence Rates) *When using the SAGA gradient estimator in Algorithm 1, set* $b\leq 4\sqrt{2}n^{2/3}$, $\gamma_k=\frac{b^3(k+\frac{4n}{b}+4)}{192n^2L}$, *and* $\tau_k=\frac{b^3}{96n^2L\gamma_k}$. *After* T *iterations, the suboptimalty at* y_T *satisfies*

$$\mathbb{E}F(y_T)-F(x^*)\leq\frac{(\frac{4n}{b}+2)(\frac{4n}{b}+4)K_1}{(T+\frac{4n}{b}+3)^2},$$

where

$$K_1=\left(F(y_0)-F(x^*)+\frac{192n^2L}{b^3(\frac{4n}{b}+2)(\frac{4n}{b}+4)}\|z_0-x^*\|^2\right).$$

If g *is* μ*-strongly convex, set* $\gamma=\min\left\{\frac{b^{3/2}}{4n\sqrt{6\mu L}},\frac{b}{4n\mu}\right\}$ *and* $\tau=\mu\gamma$. *After* T *iterations, the point* z_T *satisfies*

$$\mathbb{E}\|z_T-x^*\|^2\leq\left(1+\min\left\{\frac{b^{3/2}\sqrt{\mu}}{4n\sqrt{6L}},\frac{b}{4n}\right\}\right)^{-T}K_2,$$

where K_2 *is defined as in Theorem 6.*

It is enlightening to compare these rates to existing convergence rates for full and stochastic gradient methods. In the non-strongly convex setting, our convergence rate is $\mathcal{O}\left(n^2/T^2\right)$, matching that of Katyusha [2]. In the strongly convex case, if F is poorly conditioned so that $L/\mu \geq \mathcal{O}(b)$, we prove linear convergence at the rate $\mathcal{O}\left(\left(1+\frac{b^{3/2}\sqrt{\mu}}{n\sqrt{L}}\right)^{-T}\right)$. With $b = n^{2/3}$, this rate matches the convergence rate of inertial forward-backward on the same problem (i.e., the rate is independent of n), but we require only $n^{2/3}$ stochastic gradient evaluations per iteration compared to the n evaluations that full gradient methods require. This is reminiscent of the results of [5, 27], where the authors show that SAGA and SVRG achieve the same convergence rate as full gradient methods on non-convex problems using only $n^{2/3}$ stochastic gradient evaluations at each iteration. This is slightly worse than the results proven for Katyusha, which requires $\mathcal{O}(\sqrt{n})$ stochastic gradient evaluation per iteration to match the convergence rate of full-gradient methods.

The analogous convergence guarantees for SVRG are included in Theorem 14.

Theorem 14 (SVRG Convergence Rates) *When using the SVRG gradient estimator in Algorithm 1, set* $b \leq 32p^2$, $\gamma_k = \frac{b(k+4p+4)}{192p^2L}$, *and* $\tau_k = \frac{b}{96p^2L\gamma_k}$. *After* T *iterations, the suboptimalty at* y_T *satisfies*

$$\mathbb{E}F(y_T) - F(x^*) \leq \frac{(4p+2)(4p+4)K_1}{(T+4p+3)^2},$$

where

$$K_1 = \left(F(y_0) - F(x^*) + \frac{192p^2L}{b(4p+2)(4p+4)}\|z_0 - x^*\|^2\right).$$

If g *is* μ*-strongly convex, set* $\gamma = \min\left\{\frac{\sqrt{b}}{4p\sqrt{6\mu L}}, \frac{1}{4p\mu}\right\}$ *and* $\tau = \mu\gamma$. *After* T *iterations, the point* z_T *satisfies*

$$\mathbb{E}\|z_T - x^*\|^2 \leq \left(1 + \min\left\{\frac{\sqrt{b\mu}}{4p\sqrt{6L}}, \frac{b}{4p}\right\}\right)^{-T} K_2,$$

where K_2 *is defined as in Theorem 6.*

The convergence rates for SVRG are similar to the rates for SAGA if p and b are chosen appropriately. In the strongly convex case, setting $b = p^2$ allows SVRG to match the convergence rate of full gradient methods, and the expected number of stochastic gradient evaluations per iteration is $n/p+b$. To minimise the number of stochastic gradient evaluations while maintaining the convergence rate of full gradient methods, we set $p = \mathcal{O}(n^{1/3})$, showing that Algorithm 1 using the SVRG gradient estimator achieves the same convergence rate as full gradient methods using only $\mathcal{O}(n^{2/3})$ stochastic gradient evaluations per iteration.

The SARAH gradient estimator is similar to the SVRG estimator, as both estimators require the full gradient to be computed periodically. SARAH differs from SVRG by using previous estimates of the gradient to inform future estimates. The recursive nature of the estimator seems to decrease its MSE, which can be observed in experiments and in theory [14, 23]. However, this comes at the cost of introducing bias into the estimator.

Biased stochastic gradient methods are underdeveloped compared to their unbiased counterparts. The convergence proofs for biased algorithms are traditionally complex and difficult to generalize (see [29], for example), and proximal support has only recently been extended to biased stochastic gradient methods in the convex setting [14]. It is difficult to determine conclusively if the negative effect of the bias outweighs the benefits of a lower MSE. We show that Algorithm 1 is able to achieve an accelerated rate of convergence using biased estimators as well, beginning with the SARAH estimator.

Theorem 15 (SARAH Convergence Rates) *When using the SARAH gradient estimator in Algorithm 1, set* $b \leq 48p^4$, $\gamma_k = \frac{b(k+2p+4)}{288p^4L}$, *and* $\tau_k = \frac{b}{144p^4L\gamma_k}$. *After* T *iterations, the suboptimalty at* y_T *satisfies*

$$\mathbb{E}F(y_T) - F(x^*) \leq \frac{(2p+2)(2p+4)K_1}{(T+2p+3)^2}.$$

where

$$K_1 = \left(F(y_0) - F(x^*) + \frac{288p^4L}{b(2p+2)(2p+4)}\|z_0 - x^*\|^2\right).$$

If g is μ-strongly convex, set $\gamma = \min\left\{\sqrt{\frac{b}{144p^4\mu L}}, \frac{1}{2p\mu}\right\}$ and $\tau = \mu\gamma$. After T iterations, the point z_T satisfies

$$\mathbb{E}\|z_T - x^*\|^2 \leq \left(1 + \min\left\{\sqrt{\frac{b\mu}{144p^4 L}}, \frac{1}{2p}\right\}\right)^{-T} K_2,$$

where K_2 is defined as in Theorem 6.

We provide a proof of this result in Appendix C. Theorem 15 shows that using the SARAH gradient estimator in Algorithm 1 achieves an optimal $\mathcal{O}\left(1/T^2\right)$ convergence rate on convex objectives, but with $p = \mathcal{O}(n)$, the constant is a factor of n^2 worse than it is for accelerated SAGA, SVRG, and Katyusha. In the strongly convex case, setting $p = \mathcal{O}(n)$ and $b = \mathcal{O}(1)$ guarantees a linear convergence rate of $\mathcal{O}((1 + n^{-2}\sqrt{\mu/L})^{-T})$, achieving the optimal dependence on the condition number, but with a constant that is a factor of n worse than accelerated SAGA and SVRG, and a factor of $n^{3/2}$ worse than Katyusha. The optimal choices for b and p are $p = n^{1/5}$ and $b = p^4 = n^{4/5}$, showing that accelerated SARAH matches the convergence rate of accelerated full gradient methods using only $n^{4/5}$ stochastic gradient evaluations each iteration. Although these convergence guarantees for SARAH are slightly worse than our results for SAGA and SVRG, experimental results, including those in Section 7 and [23], show that the SARAH gradient estimator exhibits competitive performance.

Finally, we provide convergence rates for the SARGE estimator. In [14], the authors introduce the SARGE gradient estimator to mimic the recursive nature of SARAH but trade larger storage costs for a lower average per-iteration complexity, similar to the relationship between SAGA and SVRG. We prove in Appendix D that SARGE satisfies the MSEB property with similar constants to SARAH, and achieves similar convergence rates as well.

Theorem 16 (SARGE Convergence Rates) *Let[3] $c = 86016n^4/b^5$. When using the SARGE gradient estimator in Algorithm 1, set $\gamma_k = \frac{k+\frac{4n}{b}+4}{2cL}$ and $\tau_k = \frac{1}{cL\gamma_k}$. After T iterations, the suboptimalty at y_T satisfies*

$$\mathbb{E}F(y_T) - F(x^*) \leq \frac{2(\frac{2n}{b}+1)(\frac{2n}{b}+2)K_1}{(T+\frac{4n}{b}+3)^2},$$

where

$$K_1 = \left(F(y_0) - F(x^*) + \frac{86016n^4}{b^5(\frac{2n}{b}+1)(\frac{2n}{b}+2)}\|z_0 - x^*\|^2\right).$$

If g is μ-strongly convex, set $\gamma = \min\left\{\frac{1}{\sqrt{c\mu L}}, \frac{b}{4n\mu}\right\}$ and $\tau = \mu\gamma$. After T iterations, the point z_T satisfies

$$\mathbb{E}\|z_T - x^*\|^2 \leq \left(1 + \min\left\{\frac{48b^{5/2}\sqrt{154\mu}}{n^2\sqrt{L}}, \frac{b}{4n}\right\}\right)^{-T} K_2,$$

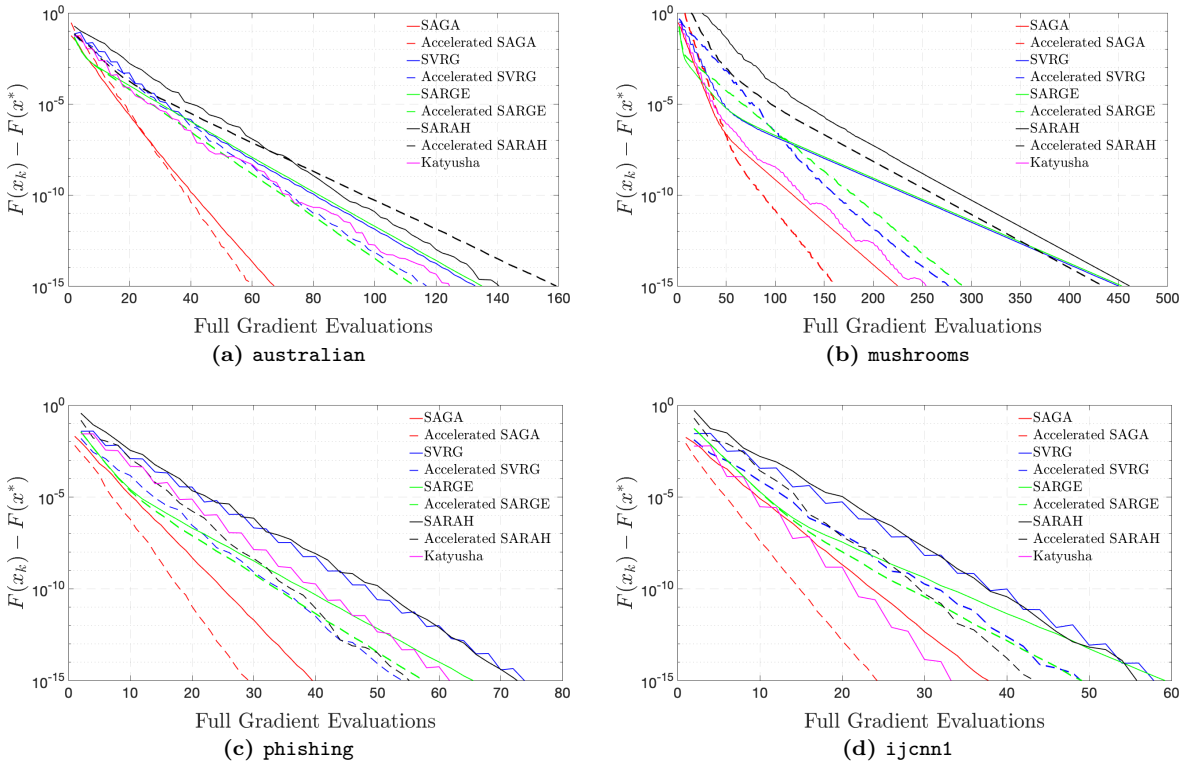
where K_2 is defined as in Theorem 6.

The convergence rates for SARGE are of the same order as the convergence rates for SARAH, even though SARGE requires fewer stochastic gradient evaluations per iteration on average. In the strongly convex case, setting $b = \mathcal{O}(n^{4/5})$ allows the algorithm to achieve the same convergence rate as full gradient methods. Even though our bound on the MSE of the SARGE estimator is a factor of n smaller than our bound on the MSE of the SAGA and SVRG estimators, the analytical difficulties due to the bias lead to a worse dependence on n. Nevertheless, SARGE is competitive in practice, as we demonstrate in the following section.

7 Numerical Experiments

To test our acceleration framework, we use it to accelerate SAGA, SVRG, SARAH, and SARGE on a series of ridge regression and LASSO tasks using the binary classification data sets `australian`, `mushrooms`,

[3]Throughout this manuscript, we have sacrificed smaller constants for generality and ease of exposition, so the constant appearing in c is not optimal.



(a) `australian` **(b)** `mushrooms`

(c) `phishing` **(d)** `ijcnn1`

Figure 1: Performance comparison for solving ridge regression among different algorithms.

`phishing`, and `ijcnn1` from the LIBSVM[4] database. We include Katyusha and Katyusha$^{\text{ns}}$ for comparison as well. For SVRG and SARAH, we compare our accelerated variants that compute the full gradient probabilistically to the non-accelerated versions that compute the full gradient deterministically at the beginning of each epoch.

With feature vectors a_i and labels y_i for $i \in \{1, 2, \cdots, n\}$, ridge regression and LASSO can be written as

$$\min_{x \in \mathbb{R}^m} \quad \frac{1}{n} \sum_{i=1}^{n} (a_i^\top x - y_i)^2 + \lambda R(x),$$

where $R \equiv \frac{1}{2}\|\cdot\|^2$ in ridge regression and $R \equiv \|\cdot\|_1$ for LASSO. Letting $g \equiv \lambda R$, it is clear that g is λ-strongly convex in ridge regression and g is not strongly convex for LASSO. In all our experiments, we rescale the value of the data to $[-1, 1]$. For ridge regression, we set $\lambda = 1/n$, and for LASSO, we set $\lambda = 1/\sqrt{n}$.

For accurate comparisons, we automate all our parameter tuning. For our experiments using ridge regression, we select the step size and momentum parameters from the set $\{1/t : t \in \mathbb{N}\}$. For LASSO, we use the parameters suggested by Theorem 5, but we scale the step size by a constant $s \in \mathbb{N}$, and we rescale the momentum parameter so that $\tau_0 = 1/2$. We perform the same parameter-tuning procedure for Katyusha, and set the negative momentum parameter $\tau_2 = 1/2$ as suggested in [2] unless otherwise stated. In our accelerated variants of SVRG and SARAH, we set $p = \frac{1}{2n}$, and for the non-accelerated variants and Katyusha, we set the epoch length to $2n$.

We measure performance with respect to the suboptimality $F(x_{k+1}) - F(x^*)$, where x^* is a low-tolerance solution found using forward-backward. To fairly compare algorithms that require a different number of stochastic gradient evaluations per iteration, we report their performance with respect to the number of effective full gradient computations they perform on average each iteration. By this metric, SAGA performs $1/n$ full gradient computations each iteration, while SVRG performs an average of $\frac{2}{n} + \frac{1}{2n}$, for example.

[4] `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

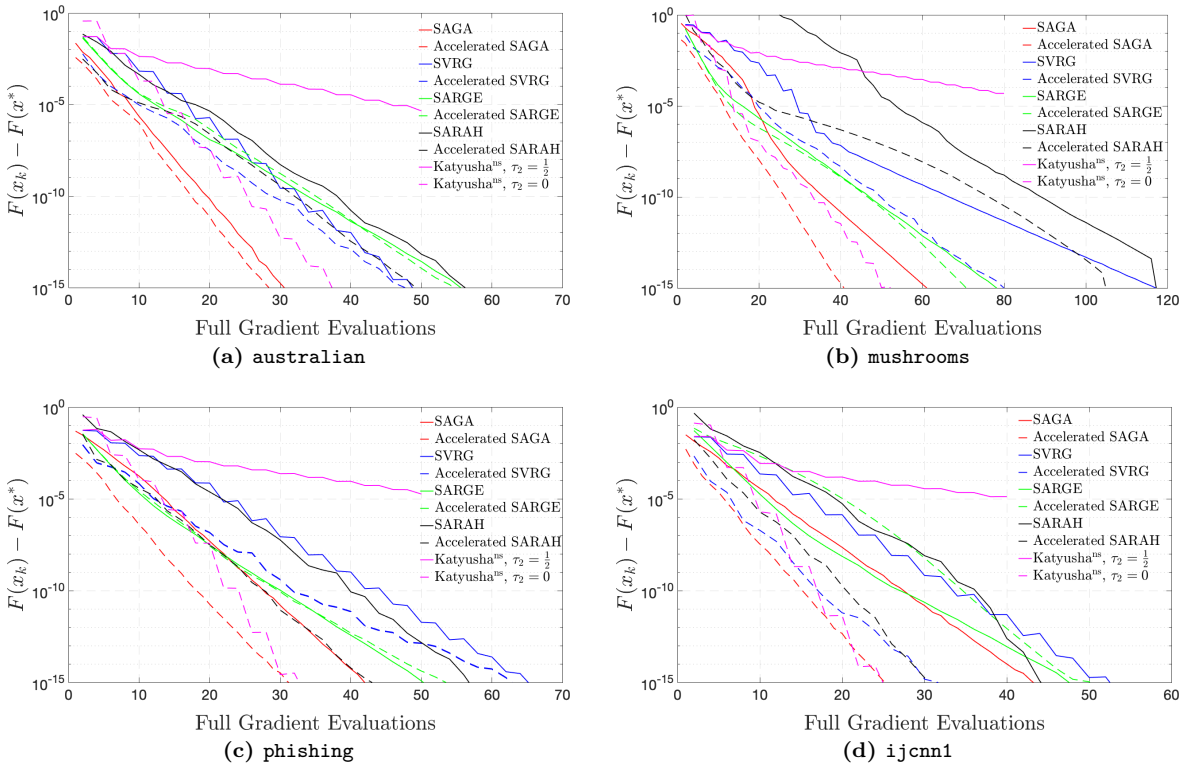


Figure 2: Performance comparison for solving LASSO among different algorithms. In Katyusha, the negative momentum parameters $\tau_2 = 0, \frac{1}{2}$ are not tuned.

Figures 1 and 2 display the median of 100 trials of ridge regression and LASSO, respectively. We observe the following trends:

- Acceleration without negative momentum significantly improves the performance of SAGA, SVRG, SARAH, and SARGE in most cases. The improvement is least dramatic on the smallest data set, `australian`, and slightly less dramatic for the biased algorithms, SARAH and SARGE.
- Because they require only one stochastic gradient evaluation per iteration, SAGA and Accelerated SAGA require significantly less computation to achieve the same accuracy as other methods.
- In the strongly convex setting, Katyusha performs similarly to or better than SVRG with acceleration in most cases.
- In the non-strongly convex setting, Katyusha$^{\text{ns}}$ performs much worse than other methods when using negative momentum. Without negative momentum, it performs much better than all algorithms except Accelerated SAGA. Because Katyusha without negative momentum is almost exactly the same algorithm as Accelerated SVRG, this improved performance is likely due to the second proximal step and additional step size η in Katyusha. All of the algorithms presented in this work can adopt these features without changing their convergence rates.

8 Conclusion

Although acceleration is a widely used and an extensively researched technique in first-order optimisation, its application to stochastic gradient methods is still poorly understood. The introduction of negative momentum adds another layer of complexity to this line of research. Although algorithms using negative momentum

enjoy fast convergence rates and strong performance when the parameters are tuned appropriately, it is unclear if negative momentum is necessary for acceleration or if it is a theoretical convenience. In this work, we propose a universal framework for accelerating stochastic gradient methods that does not rely on negative momentum.

Because our approach does not rely on negative momentum, it applies to a much broader class of stochastic gradient estimators. As long as the estimator admits natural bounds on its bias and MSE, it can be used in our framework to produce an accelerated stochastic gradient method with an optimal $1/T^2$ dependence on convex problems and an optimal $\sqrt{\kappa}$ dependence in the strongly convex setting. The bias and MSE of the estimator appear only in the constants of our convergence rates. From this perspective, negative momentum is effectively a variance-reduction technique, reducing the variance in the iterates to improve the dependence on n in the convergence rates. A natural question for future research is whether there exist gradient estimators with smaller bias and MSE than SAGA, SVRG, SARAH, and SARGE that can be accelerated using our framework and admit a better dependence on n.

References

[1] Allen-Zhu, Z. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *ICML* (2017).

[2] Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research 18* (2018), 1–51.

[3] Allen-Zhu, Z. Katyusha X: Practical momentum method for stochastic sum-of-nonconvex optimization. In *ICML* (2018).

[4] Allen-Zhu, Z. Natasha 2: Faster non-convex optimization than SGD. In *Advances in Neural Information Processing Systems* (2018).

[5] Allen-Zhu, Z., and Hazan, E. Variance reduction for faster non-convex optimization. In *Proceedings of the* 33^{rd} *International Conference on Machine Learning* (2016), vol. 48.

[6] Allen-Zhu, Z., and Orecchia, L. Linear coupling: An ultimate unification of gradient and mirror descent. In *Proceedings of the* 8^{th} *Innovations in Theoretical Computer Science (ITCS)* (2017).

[7] Beck, A., and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*, 1 (2009), 183–202.

[8] Bottou, L., Curtis, F. E., , and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review 60* (2018), 223–311.

[9] Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM* (2009).

[10] Candés, E. J., and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* (2009), 717–772.

[11] Combettes, P. L., and Wajs, V. R. Signal recovery by proximal forward-backward splitting. *Multiscale Modelling and Simulation 4*, 4 (2005), 1168–1200.

[12] Defazio, A. A simple practical accelerated method for finite sums. In *Advances In Neural Information Processing Systems* (2016), pp. 676–684.

[13] Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems* (2014), pp. 1646–1654.

[14] Driggs, D., Liang, J., and Schönlieb, C.-B. A unified analysis of biased stochastic gradient methods–and one new one. *Technical Report, University of Cambridge* (2019).

[15] Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In 32^{nd} *Conference on Neural Information Processing Systems* (2018).

[16] Frostig, R., Ge, R., Kakade, S. M., and Sidford, A. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *ICML* (2015), vol. 37, pp. 1–28.

[17] Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems* (2015).

[18] Johnson, R., and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems* (2013), pp. 315–323.

[19] Lan, G., Li, Z., and Zhou, Y. A unified variance-reduced accelerated gradient method for convex optimization. *arXiv:1905.12412* (2019).

[20] Lin, H., Mairal, J., and Harchaoui, Z. A universal catalyst for first-order optimization. In *Advances In Neural Information Processing Systems* (2015).

[21] Lustig, M., Donoho, D., and Pauly, J. M. SparseMRI: the application of compressed sensing for rapid MR imaging. *Magnetic Resonance Medicine 6* (2007), 1182–1195.

[22] Nesterov, Y. *Introductory lectures on convex programming.* Springer, 2004.

[23] Nguyen, L. M., Liu, J., Scheinberg, K., and Takáĉ, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning* (2017), vol. 70, pp. 2613–2621.

[24] Nitanda, A. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems* (2014), pp. 1574–1582.

[25] Passty, G. B. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications 72*, 2 (1979), 383–390.

[26] Pham, N. H., Nguyen, L. M., Phan, D. T., and Tran-Dinh, Q. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *arXiv:1902.05679* (2019).

[27] Reddi, S. J., Sra, S., Póczós, B., and Smola, A. Fast stochastic methods for nonsmooth nonconvex optimization. In *ICML* (2016).

[28] Robbins, H., and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics 22*, 3 (1951), 400–407.

[29] Schmidt, M., Roux, N. L., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming 162* (2017), 83–112.

[30] Shang, F., Liu, Y., Cheng, J., and Zhuo, J. Fast stochastic variance reduced gradient method with momentum acceleration for machine learning. In *Proceedings of the* 34^{th} *International Conference on Machine Learning (ICML)* (2017).

[31] Tibshirani, R. Regression shrinkage and variable selection via the lasso. *Journal of the Royal Statistical Society, Series B* (1996), 267–288.

[32] Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. SpiderBoost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv:1810.10690* (2018).

[33] Woodworth, B., and Srebro, N. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems* (2016).

[34] Xiao, L., and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization 24*, 4 (2014), 2057–2075.

[35] Zhang, Y., and Xiao, L. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *ICML* (2015).

[36] Zhou, K. Direct acceleration of saga using sampled negative momentum. *arXiv:1806.11048* (2018).

[37] Zhou, K., Shang, F., and Cheng, J. A simple stochastic variance reduced algorithm with fast convergence rates. In *ICML* (2018).

[38] Zhou, Y., Wang, Z., Ji, K., Liang, Y., and Tarokh, V. Momentum schemes with stochastic variance reduction for nonconvex composite optimization. *arXiv:1902.02715* (2019).

A Proofs of Non-Positivity

The goal is to show that the two terms

$$\frac{\rho_M(1-\rho_B)}{8\tau_k^2} + 4\gamma^2 L^2 \Theta_1 \Theta_2 + \frac{\gamma_k}{\tau_k}\left(\frac{L}{2} - \frac{1}{4\tau_k\gamma_k}\right) \quad \text{and} \quad 8\gamma_k^2 L \Theta_1 \Theta_2 - \frac{\gamma_k(1-\tau_k)}{\tau_k} \tag{14}$$

are non-positive with the parameter choices of Theorems 5 and 6. We consider three cases.

Case 1. Let γ_k and τ_k be as in the statement of Theorem 5. For the first term in (14),

$$\begin{aligned}
&\frac{\rho_M(1-\rho_B)}{8\tau_k^2} + 4\gamma_k^2 L^2 \Theta_1 \Theta_2 + \frac{\gamma_k}{\tau_k}\left(\frac{L}{2} - \frac{1}{4\tau_k\gamma_k}\right) \\
=&\gamma_k^2 L^2 \left(\frac{\rho_M(1-\rho_B)c^2}{8} + 4\Theta_1\Theta_2 + 4c\left(\frac{1}{2} - \frac{c}{4}\right)\right)
\end{aligned}$$

The constraint

$$c \geq \frac{2}{2 - \rho_M + \rho_B\rho_M}\left(1 + \sqrt{1 + 8\Theta_1\Theta_2(2 - \rho_M + \rho_B\rho_M}\right)$$

ensures that this quadratic in c is non-positive. For the second term, we require $\tau_k \leq 1/2$ for all k, which holds because $\tau_k = \frac{2}{k+\nu+4} \leq \frac{1}{2}$. Therefore,

$$8\gamma_k^2 L \Theta_1 \Theta_2 - \frac{\gamma_k(1-\tau_k)}{\tau_k} \leq 8\gamma_k^2 L \Theta_1 \Theta_2 - \frac{cL\gamma_k^2 \Theta_1 \Theta_2}{2}.$$

The constraint $c \geq 16\Theta_1\Theta_2$ implies that this quantity is non-positive.

Case 2. Let γ and τ be as in the statement of Theorem 6, and suppose $\frac{1}{\sqrt{\mu L c}} \leq \frac{\rho}{2\mu}$. In this case, $\tau = \sqrt{\frac{\mu}{Lc}} = \frac{1}{Lc\gamma}$. As in Case 1,

$$\begin{aligned}
&\frac{\rho_M(1-\rho_B)}{8\tau^2} + 4\gamma^2 L^2 \Theta_1 \Theta_2 + \frac{\gamma}{\tau}\left(\frac{L}{2} - \frac{1}{4\tau\gamma}\right) \\
=&\gamma^2 L^2 \left(\frac{\rho_M(1-\rho_B)c^2}{8} + 4\Theta_1\Theta_2 + 4c\left(\frac{1}{2} - \frac{c}{4}\right)\right),
\end{aligned}$$

which is non-positive due to the constraints on c. For the second term, all we must show is that $1 - \tau \geq 1/2$. We have $\tau = \sqrt{\frac{\mu}{Lc}} \leq \frac{1}{\sqrt{c}}$, and c is larger than 4, so the constraint $c \geq 16\Theta_1\Theta_2$ ensures that the second term in (14) is non-positive.

Case 3. In Theorem 6, suppose instead that $\frac{\rho}{2\mu} \leq \frac{1}{\sqrt{\mu L c}}$, so that $\gamma = \frac{\rho}{2\mu}$ and $\tau = \frac{\rho}{2}$. This assumption implies the inequality $\frac{L}{\mu} \leq \frac{4}{c\rho^2}$, so

$$\begin{aligned}
&\frac{\rho_M(1-\rho_B)}{8\tau^2} + 4\gamma^2 L^2 \Theta_1 \Theta_2 + \frac{\gamma}{\tau}\left(\frac{L}{2} - \frac{1}{4\tau\gamma}\right) \\
&= \frac{\rho_M(1-\rho_B)}{8\mu^2\gamma^2} + 4\gamma^2 L^2 \Theta_1 \Theta_2 + \frac{1}{\mu}\left(\frac{L}{2} - \frac{1}{4\mu\gamma^2}\right) \\
&= \frac{\rho_M(1-\rho_B)}{2\rho^2} + \frac{\rho^2 L^2 \Theta_1 \Theta_2}{\mu^2} + \frac{L}{2\mu} - \frac{1}{\rho^2} \\
&\leq \frac{\rho_M(1-\rho_B)}{2\rho^2} + \frac{16\Theta_1\Theta_2}{c^2\rho^2} + \frac{2}{c\rho^2} - \frac{1}{\rho^2} \\
&= \frac{1}{c^2\rho^2}\left(\frac{\rho_M(1-\rho_B)c^2}{2} + 16\Theta_1\Theta_2 + 2c - c^2\right).
\end{aligned}$$

This is a quadratic in c with the root

$$\frac{2}{2-\rho_M+\rho_B\rho_M}\left(1+\sqrt{1+8\Theta_1\Theta_2(2-\rho_M+\rho_B\rho_M}\right).$$

Because c is larger than this quantity, this term is non-positive. For the second term in (14),

$$8\gamma^2 L\Theta_1\Theta_2-\frac{\gamma(1-\tau)}{\tau}=\frac{2L\Theta_1\Theta_2\rho^2}{\mu^2}-\frac{1}{2\mu}\leq\frac{8\Theta_1\Theta_2}{c\mu}-\frac{1}{2\mu}\leq 0,$$

where the last inequality follows from the fact that $c\geq 16\Theta_1\Theta_2$.

B Proofs for SAGA and SVRG

We begin with a standard bound on the variance $\|\widetilde{\nabla}^{\text{SAGA}}_{k+1}-\nabla f(x_{k+1})\|^2$ that is an easy consequence of the variance bound in [13], but [13] and related works [2, 12, 18, 34] ultimately use a much looser bound in their convergence analysis.

Lemma 17 *The variance of the SAGA gradient estimator with minibatches of size b is bounded as follows:*

$$\mathbb{E}_k\|\widetilde{\nabla}^{\text{SAGA}}_{k+1}-\nabla f(x_{k+1})\|^2\leq\frac{1}{bn}\sum_{i=1}^n\mathbb{E}_k\|\nabla f_i(x_{k+1})-\nabla f_i(\varphi_k^i)\|^2$$

Proof. Recall that for any random variable X, $\arg\min_Y\mathbb{E}\|X-Y\|^2=\mathbb{E}X$. With $X=\frac{1}{b}\sum_{j\in J_k}\nabla f_j(x_{k+1})-\nabla f_j(\varphi_k^j)$, this implies

$$\begin{aligned}
&\mathbb{E}_k\|\widetilde{\nabla}^{\text{SAGA}}_{k+1}-\nabla f(x_{k+1})\|^2\\
=&\mathbb{E}_k\|\nabla f_j(x_{k+1})-\nabla f_j(\varphi_k^j)+\frac{1}{n}\sum_{i=1}^n\nabla f_i(\varphi_k^i)-\nabla f(x_{k+1})\|^2\\
=&\mathbb{E}_k\|X-\mathbb{E}_kX\|^2\\
\leq&\mathbb{E}_k\|X\|^2\\
=&\frac{1}{b^2}\mathbb{E}_k\left\|\sum_{j\in J_k}\nabla f_j(x_{k+1})-\nabla f_j(\varphi_k^j)\right\|^2\\
\overset{①}{\leq}&\frac{1}{b^2}\mathbb{E}_k\sum_{j\in J_k}\|\nabla f_j(x_{k+1})-\nabla f_j(\varphi_k^j)\|^2\\
\overset{②}{=}&\frac{1}{bn}\sum_{i=1}^n\|\nabla f_i(x_{k+1})-\nabla f_i(\varphi_k^i)\|^2.
\end{aligned}$$

Inequality ① is Jensen's, and ② comes from computing the expectation. □

Lemma 17 provides a variance bound that is compatible with the MSEB property, as we show in the following lemma.

Lemma 18 *The SAGA gradient estimator satisfies the MSEB property with $M_1=\frac{3n}{b^2}$, $\rho_M=\frac{b}{2n}$, $M_2=0$, and $\rho_B=\rho_F=1$.*

Proof. Lemma 17 shows that the MSE of the SAGA gradient estimator is dominated by $\frac{1}{bn}\sum_{i=1}^n\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(\varphi_k^i)\|^2$, so we choose this sequence for $\mathcal{M}_k$. Using the inequality $\|a-c\|^2\leq(1+\frac{2n}{b})\|a-b\|^2+(1+\frac{b}{2n})\|b-c\|^2$,

$$\mathcal{M}_k=\frac{1}{bn}\sum_{i=1}^n\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(\varphi_k^i)\|^2$$

$$
\begin{aligned}
&\leq \frac{1+\frac{2n}{b}}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2+\frac{1+\frac{b}{2n}}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_k)-\nabla f_i(\varphi_k^i)\|^2\\
&\overset{①}{=} \frac{1+\frac{2n}{b}}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2+\frac{1+\frac{b}{2n}}{bn}\left(1-\frac{b}{n}\right)\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_k)-\nabla f_i(\varphi_{k-1}^i)\|^2\\
&\overset{②}{\leq} \frac{3}{b^2}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2+\frac{1}{bn}\left(1-\frac{b}{2n}\right)\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_k)-\nabla f_i(\varphi_{k-1}^i)\|^2\\
&= \frac{3}{b^2}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2+\left(1-\frac{b}{2n}\right)\mathcal{M}_{k-1}.
\end{aligned}
$$

Equality ① follows from computing expectations and the update rule for φ_k^i:

$$
\begin{aligned}
\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_k)-\nabla f_i(\varphi_k^i)\|^2 &= \frac{1}{b}\sum_{j\in J_{k-1}}\mathbb{E}\|\nabla f_j(x_k)-\nabla f_j(\varphi_k^j)\|^2\\
&\quad+\mathbb{E}\sum_{i\notin J_{k-1}}^{n}\|\nabla f_i(x_k)-\nabla f_i(\varphi_{k-1}^i)\|^2\\
&= 0+\left(1-\frac{b}{n}\right)\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_k)-\nabla f_i(\varphi_{k-1}^i)\|^2,
\end{aligned}
$$

and ② follows from the the inequalities $\left(1+\frac{b}{2n}\right)\left(1-\frac{b}{n}\right)\leq\left(1-\frac{b}{2n}\right)$ and $1+\frac{2n}{b}\leq\frac{3n}{b}$. This shows that we can take $M_1=\frac{3n}{b^2}$, $M_2=0$, and $\rho_F=1$. Because the SAGA gradient estimator is unbiased, we can clearly set $\rho_B=1$, proving the claim. □

A similar result holds for the SVRG gradient estimator.

Corollary 19 *The SVRG gradient estimator satisfies the MSEB property with $M_1=\frac{3p}{b}$, $\rho_M=\frac{1}{2p}$, $M_2=0$, and $\rho_B=\rho_F=1$.*

Proof. Following the same argument as in the proof of Lemma 17, we have the bound

$$
\mathbb{E}\|\widetilde{\nabla}_{k+1}^{\mathrm{SVRG}}-\nabla f(x_{k+1})\|^2\leq\frac{1-1/p}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(\widetilde{x})\|^2.
$$

The factor $1-1/p$ that appears is due to the fact that $\widetilde{\nabla}_{k+1}=\nabla f(x_{k+1})$ with probability $1/p$. With $\mathcal{M}_k=\frac{1-1/p}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(\widetilde{x})\|^2$, we follow the proof of Lemma 18.

$$
\begin{aligned}
\mathcal{M}_k &= \frac{1-1/p}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(\widetilde{x})\|^2\\
&\leq \frac{(1+2p)(1-1/p)}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2+\frac{(1+\frac{1}{2p})(1-1/p)}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_k)-\nabla f_i(\widetilde{x})\|^2\\
&\overset{①}{=} \frac{(1+2p)(1-1/p)}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2+\frac{(1+\frac{1}{2p})(1-1/p)^2}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_k)-\nabla f_i(\widetilde{x})\|^2\\
&\leq \frac{3p}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2+\left(1-\frac{1}{2p}\right)\mathcal{M}_{k-1}.
\end{aligned}
$$

Equality ① follows from the fact that $\widetilde{x}=x_k$ with probability $1/p$. □

With the MSEB property established for the SAGA and SVRG gradient estimators, we can apply Theorems 5 and 6 to get a rate of convergence. For the SAGA estimator, Lemma 18 ensures that the choices $c=\frac{96n^2}{b^3}$ and $\rho=\frac{b}{2n}$ satisfy the hypotheses of Theorems 5 and 6 as long as $b\leq 4\sqrt{2}n^{2/3}$. Similarly, for the SVRG estimator, the choices $c=\frac{b}{96p^2}$ and $\rho=\frac{1}{2p}$ satisfy the conditions of Theorems 5 and 6 as long as $b\leq 32p^2$.

C Proofs for SARAH

To prove the convergence rates of Theorem 15, we first show that the SARAH gradient estimator satisfies the MSEB property.

Lemma 20 *The SARAH gradient estimator satisfies the MSEB property with $M_1 = 1/b$, $M_2 = 0$, $\rho_M = 1/p$, $\rho_B = 1/p$, and $\rho_F = 1$.*

Proof. The SARAH gradient estimator is equal to $\nabla f(x_{k+1})$ with probability $1/p$, so the expectation of the SARAH gradient estimator is

$$\begin{aligned}\mathbb{E}_k \widetilde{\nabla}_{k+1}^{\text{SARAH}} &= \frac{1}{p}\nabla f(x_{k+1}) + \frac{1}{b}\mathbb{E}_k\left(\sum_{j\in J_k} \nabla f_j(x_{k+1}) - f_j(x_k)\right) + \widetilde{\nabla}_k^{\text{SARAH}} \\ &= \frac{1}{p}\nabla f(x_{k+1}) + \left(1-\frac{1}{p}\right)\left(\nabla f(x_{k+1}) - \nabla f(x_k) + \widetilde{\nabla}_k^{\text{SARAH}}\right)\end{aligned}$$

Therefore,

$$\nabla f(x_{k+1}) - \mathbb{E}_k \widetilde{\nabla}_{k+1}^{\text{SARAH}} = \left(1-\frac{1}{p}\right)\left(\nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARAH}}\right),$$

so $\rho_B = 1/p$. Next, we prove a bound on the MSE. The beginning of our proof is similar to the proof of the MSE bound in [23, Lem. 2].

$$\begin{aligned}&\mathbb{E}_k\|\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \nabla f(x_{k+1})\|^2 \\ =&\mathbb{E}_k\left\|\widetilde{\nabla}_k^{\text{SARAH}} - \nabla f(x_k) + \nabla f(x_k) - \nabla f(x_{k+1}) + \widetilde{\nabla}_{k+1}^{\text{SARAH}} - \widetilde{\nabla}_k^{\text{SARAH}}\right\|^2 \\ =&\left\|\widetilde{\nabla}_k^{\text{SARAH}} - \nabla f(x_k)\right\|^2 + \|\nabla f(x_k) - \nabla f(x_{k+1})\|^2 + \mathbb{E}_k\left\|\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \widetilde{\nabla}_k^{\text{SARAH}}\right\|^2 \\ &+ 2\langle \nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARAH}}, \nabla f(x_{k+1}) - \nabla f(x_k)\rangle \\ &- 2\left\langle \nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARAH}}, \mathbb{E}_k\left[\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \widetilde{\nabla}_k^{\text{SARAH}}\right]\right\rangle \\ &- 2\left\langle \nabla f(x_{k+1}) - \nabla f(x_k), \mathbb{E}_k\left[\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \widetilde{\nabla}_k^{\text{SARAH}}\right]\right\rangle.\end{aligned}$$

We consider each inner product separately. The first inner product is equal to

$$\begin{aligned}&2\langle \nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARAH}}, \nabla f(x_{k+1}) - \nabla f(x_k)\rangle \\ =&-\|\nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARAH}}\|^2 - \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \\ &+ \|\nabla f(x_{k+1}) - \widetilde{\nabla}_k^{\text{SARAH}}\|^2.\end{aligned}$$

For the next two inner products, we use the fact that

$$\mathbb{E}_k[\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \widetilde{\nabla}_k^{\text{SARAH}}] = \nabla f(x_{k+1}) - \nabla f(x_k).$$

With this equality established, we see that the second inner product is equal to

$$\begin{aligned}&-2\left\langle \nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARAH}}, \mathbb{E}_k\left[\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \widetilde{\nabla}_k^{\text{SARAH}}\right]\right\rangle \\ =&-2\langle \nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARAH}}, \nabla f(x_{k+1}) - \nabla f(x_k)\rangle \\ =&\|\nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARAH}}\|^2 + \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 - \|\nabla f(x_{k+1}) - \widetilde{\nabla}_k^{\text{SARAH}}\|^2.\end{aligned}$$

The third inner product can be bounded using a similar procedure.

$$-2\left\langle \nabla f(x_{k+1}) - \nabla f(x_k), \mathbb{E}_k\left[\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \widetilde{\nabla}_k^{\text{SARAH}}\right]\right\rangle$$

$$
\begin{aligned}
&= -2\langle \nabla f(x_{k+1}) - \nabla f(x_k), \nabla f(x_{k+1}) - \nabla f(x_k)\rangle \\
&= -2\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2.
\end{aligned}
$$

Altogether and after applying the full expectation operator, we have

$$
\begin{aligned}
&\mathbb{E}\|\widetilde{\nabla}_k^{\text{SARAH}} - \nabla f(x_{k+1})\|^2 \\
\leq &\mathbb{E}\left\|\widetilde{\nabla}_k^{\text{SARAH}} - \nabla f(x_k)\right\|^2 - \mathbb{E}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \\
&+ \mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \widetilde{\nabla}_k^{\text{SARAH}}\|^2 \\
\leq &\mathbb{E}\left\|\widetilde{\nabla}_k^{\text{SARAH}} - \nabla f(x_k)\right\|^2 + \mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \widetilde{\nabla}_k^{\text{SARAH}}\|^2.
\end{aligned}
$$

We can simplify the final line by computing expectations. With probability $1/p$, $\widetilde{\nabla}_k^{\text{SARAH}} = \nabla f(x_k)$, so

$$
\mathbb{E}\left\|\widetilde{\nabla}_k^{\text{SARAH}} - \nabla f(x_k)\right\|^2 = \left(1 - \frac{1}{p}\right)\mathbb{E}\left\|\widetilde{\nabla}_k^{\text{SARAH}} - \nabla f(x_k)\right\|^2.
$$

For the second term,

$$
\begin{aligned}
\mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \widetilde{\nabla}_k^{\text{SARAH}}\|^2 =&\mathbb{E}\left\|\frac{1}{b}\left(\sum_{j\in J_k}\nabla f_j(x_{k+1}) - \nabla f_j(x_k)\right)\right\|^2 \\
\leq&\frac{1}{b^2}\mathbb{E}\left[\sum_{j\in J_k}\|\nabla f_j(x_{k+1}) - \nabla f_j(x_k)\|^2\right] \\
=&\frac{1}{bn}\sum_{i=1}^n \mathbb{E}\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|^2.
\end{aligned}
$$

The inequality is Jensen's. This results in the recursive inequality

$$
\begin{aligned}
&\mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \nabla f(x_{k+1})\|^2 \\
\leq &\left(1 - \frac{1}{p}\right)\mathbb{E}\left\|\widetilde{\nabla}_k^{\text{SARAH}} - \nabla f(x_k)\right\|^2 + \frac{1}{bn}\sum_{i=1}^n \mathbb{E}\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|^2.
\end{aligned}
$$

With $\mathcal{M}_k = \mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARAH}} - \nabla f(x_{k+1})\|^2$, it is clear that we can take $M_1 = 1/b, \rho_M = 1/p$, $M_2 = 0$, and $\rho_F = 1$. □

With these MSEB constants established convergence rates easily follow from Theorems 5 and 6 with $c = 144p^4/b$ and $\rho = 1/p$.

D Proofs for SARGE

For the proofs in this section, we rewrite the SARGE gradient estimator in terms of the SAGA estimator to make the analysis easier to follow. Define the operator

$$
\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}} \overset{\text{def}}{=} \frac{1}{b}\left(\sum_{j\in J_k}\nabla f_j(x_k) - \nabla f_j(\xi^j)\right) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\xi^i),
$$

where the variables $\{\xi_k^i\}_{i=1}^n$ follow the update rules $\xi_{k+1}^j = x_k$ and $\xi_{k+1}^i = \xi_k^i$ for all $i \notin J_k$. The SARGE estimator is equal to

$$
\widetilde{\nabla}_{k+1}^{\text{SARGE}} = \widetilde{\nabla}_{k+1}^{\text{SAGA}} - \left(1 - \frac{b}{n}\right)\left(\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}} - \widetilde{\nabla}_{k+1}^{\text{SARGE}}\right).
$$

Before we begin, we require a bound on the MSE of the ξ-SAGA gradient estimator that follows immediately from Lemma 18.

Lemma 21 *The MSE of the ξ-SAGA gradient estimator satisfies the following bound:*

$$\mathbb{E}\left\|\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}}-\nabla f(x_k)\right\|^2 \leq \frac{3}{b^2}\sum_{\ell=1}^{k}\left(1-\frac{b}{2n}\right)^{k-\ell}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_\ell)-\nabla f_i(x_{\ell-1})\|^2.$$

Proof. Following the proof of Lemma 17,

$$\begin{aligned}
&\mathbb{E}_k\left\|\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}}-\nabla f(x_k)\right\|^2\\
=&\mathbb{E}_k\left\|\frac{1}{b}\sum_{j\in J_k}\left(\nabla f_j(x_k)-\nabla f_j(\xi_k^j)\right)-\nabla f(x_k)+\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\xi_k^i)\right\|^2\\
\overset{①}{=}&\frac{1}{bn}\sum_{i=1}^{n}\left\|\nabla f_i(x_k)-\nabla f_i(\xi_k^i)\right\|^2-\left\|\nabla f(x_k)-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\xi_k^i)\right\|^2\\
\leq&\frac{1}{bn}\sum_{i=1}^{n}\left\|\nabla f_i(x_k)-\nabla f_i(\xi_k^i)\right\|^2.
\end{aligned}$$

Equality ① is the standard variance decomposition. To continue, we follow the proof of Lemma 18.

$$\begin{aligned}
&\mathbb{E}\left\|\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}}-\nabla f(x_k)\right\|^2\\
\leq&\frac{1}{bn}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(x_k)-\nabla f_i(\xi_k^i)\right\|^2\\
\leq&\frac{(1+\frac{2n}{b})}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_k)-\nabla f_i(x_{k-1})\|^2+\frac{1}{bn}\left(1+\frac{b}{2n}\right)\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(x_{k-1})-\nabla f_i(\xi_k^i)\right\|^2\\
\overset{②}{=}&\frac{(1+\frac{2n}{b})}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_k)-\nabla f_i(x_{k-2})\|^2+\frac{1}{bn}\left(1+\frac{b}{2n}\right)\left(1-\frac{b}{n}\right)\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(x_{k-1})-\nabla f_i(\xi_{k-1}^i)\right\|^2\\
\overset{③}{\leq}&\frac{3}{b^2}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_k)-\nabla f_i(x_{k-1})\|^2+\frac{1}{bn}\left(1-\frac{b}{2n}\right)\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(x_{k-1})-\nabla f_i(\xi_{k-1}^i)\right\|^2\\
\leq&\frac{3}{b^2}\sum_{\ell=1}^{k}\left(1-\frac{b}{2n}\right)^{k-\ell}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_\ell)-\nabla f_i(x_{\ell-1})\|^2.
\end{aligned}$$

Equality ② follows from computing expectations, and ③ uses the estimate $\left(1-\frac{b}{n}\right)\left(1+\frac{b}{2n}\right)\leq\left(1-\frac{b}{2n}\right)$. □

Due to the recursive nature of the SARGE gradient estimator, its MSE depends on the difference between the current estimate and the estimate from the previous iteration. This is true for the recursive SARAH gradient estimate as well, but bounding the quantity $\|\widetilde{\nabla}_k^{\text{SARAH}}-\widetilde{\nabla}_{k-1}^{\text{SARAH}}\|^2$ is a much more straightforward task than bounding the same quantity for the SARGE estimator. The next lemma provides this bound.

Lemma 22 *The SARGE gradient estimator satisfies the following bound:*

$$\begin{aligned}
&\mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARGE}}-\widetilde{\nabla}_k^{\text{SARGE}}\|^2\\
\leq&\frac{27+12b}{n^2b^2}\sum_{\ell=2}^{k}\left(1-\frac{b}{2n}\right)^{k-\ell}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{\ell-1})-\nabla f_i(x_{\ell-2})\|^2\\
&+\frac{12}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2+\frac{3}{2n^2}\mathbb{E}\left\|\nabla f(x_k)-\widetilde{\nabla}_k^{\text{SARGE}}\right\|^2.
\end{aligned}$$

Proof. To begin, we use the standard inequality $\|a-c\|^2\leq(1+\delta)\|a-b\|^2+(1+\delta^{-1})\|b-c\|^2$ for any $\delta>0$ twice. For simplicity, we set $\delta=\sqrt{3/2}-1$ and use the fact that $1+\frac{1}{\sqrt{3/2}-1}\leq 6$ for both applications of

this inequality.

$$
\begin{aligned}
&\mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARGE}} - \widetilde{\nabla}_{k}^{\text{SARGE}}\|^2 \\
=&\mathbb{E}\left\|\widetilde{\nabla}_{k+1}^{\text{SAGA}} - \left(1-\frac{b}{n}\right)\left(\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}} - \widetilde{\nabla}_{k}^{\text{SARGE}}\right) - \widetilde{\nabla}_{k}^{\text{SARGE}}\right\|^2 \\
\leq&6\mathbb{E}\left\|\widetilde{\nabla}_{k+1}^{\text{SAGA}} - \widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}}\right\|^2 + \frac{\sqrt{3}b^2}{\sqrt{2}n^2}\mathbb{E}\left\|\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}} - \widetilde{\nabla}_{k}^{\text{SARGE}}\right\|^2 \\
\leq&6\mathbb{E}\left\|\widetilde{\nabla}_{k+1}^{\text{SAGA}} - \widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}}\right\|^2 + \frac{6\sqrt{3}b^2}{\sqrt{2}n^2}\mathbb{E}\left\|\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}} - \nabla f(x_k)\right\|^2 + \frac{3b^2}{2n^2}\mathbb{E}\left\|\nabla f(x_k) - \widetilde{\nabla}_{k}^{\text{SARGE}}\right\|^2 \\
\leq&6\mathbb{E}\left\|\widetilde{\nabla}_{k+1}^{\text{SAGA}} - \widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}}\right\|^2 + \frac{9b^2}{n^2}\mathbb{E}\left\|\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}} - \nabla f(x_k)\right\|^2 + \frac{3b^2}{2n^2}\mathbb{E}\left\|\nabla f(x_k) - \widetilde{\nabla}_{k}^{\text{SARGE}}\right\|^2.
\end{aligned} \tag{15}
$$

We now bound the first two of these three terms separately. Consider the first term.

$$
\begin{aligned}
&6\mathbb{E}\left\|\widetilde{\nabla}_{k+1}^{\text{SAGA}} - \widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}}\right\|^2 \\
=&6\mathbb{E}\left\|\frac{1}{b}\left(\sum_{j\in J_k}\nabla f_j(x_{k+1}) - \nabla f_j(\varphi_k^j)\right) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i)\right. \\
&\left. - \frac{1}{b}\left(\sum_{j\in J_{k-1}}\nabla f_j(x_k) - \nabla f_j(\xi_k^j)\right) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\xi_k^i)\right\|^2 \\
\leq&12\mathbb{E}\left\|\frac{1}{b}\left(\sum_{j\in J_k}\nabla f_j(x_{k+1}) - \nabla f_j(x_k)\right)\right\|^2 \\
&+ 12\mathbb{E}\left\|\frac{1}{b}\left(\sum_{j\in J_k}\nabla f_j(\varphi_k^j) - \nabla f_j(\xi_k^j)\right) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\xi_k^i)\right\|^2 \\
\overset{①}{=}&12\mathbb{E}\left\|\frac{1}{b}\left(\sum_{j\in J_k}\nabla f_j(x_{k+1}) - \nabla f_j(x_k)\right)\right\|^2 + 12\mathbb{E}\left\|\frac{1}{b}\left(\sum_{j\in J_k}\nabla f_j(\varphi_k^j) - \nabla f_j(\xi_k^j)\right)\right\|^2 \\
&- 12\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\varphi_k^i) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\xi_k^i)\right\|^2 \\
\leq&\frac{12}{bn}\sum_{i=1}^n \mathbb{E}\left\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\right\|^2 + 12\mathbb{E}\left\|\frac{1}{b}\left(\sum_{j\in J_k}\nabla f_j(\varphi_k^j) - \nabla f_j(\xi_k^j)\right)\right\|^2 \\
\leq&\frac{12}{bn}\sum_{i=1}^n \mathbb{E}\left\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\right\|^2 + \frac{12}{b^2}\mathbb{E}\sum_{j\in J_k}\left\|\nabla f_j(\varphi_k^j) - \nabla f_j(\xi_k^j)\right\|^2.
\end{aligned}
$$

Equality ① is the standard variance decomposition, which states that for any random variable X, $\mathbb{E}_k\|X - \mathbb{E}_k X\|^2 = \mathbb{E}_k\|X\|^2 - \|\mathbb{E}_k X\|^2$. The second term can be reduced further by computing the expectation. Let j_k be any element of J_k. The probability that $\nabla f_{j_k}(\varphi_k^{j_k}) = \nabla f_{j_{k-1}}(x_{k+1})$ is equal to the probability that $j_k = j_{k-1}$, which is $1/n$. The probability that $\nabla f_{j_k}(\varphi_k^{j_k}) = \nabla f_{j_{k-2}}(x_k)$ is equal to the probability that $j_k \neq j_{k-1}$ and $j = j_{k-2}$, which is $1/n\,(1 - b/n)$. Continuing in this way,

$$
\mathbb{E}\left\|\nabla f_{j_k}(\varphi_k^{j_k}) - \nabla f_{j_k}(\xi_k^{j_k})\right\|^2 = \frac{1}{n}\sum_{\ell=1}^k \left(1-\frac{b}{n}\right)^{k-\ell}\mathbb{E}\|\nabla f_{j_{\ell-1}}(x_{\ell+1}) - \nabla f_{j_{\ell-1}}(x_\ell)\|^2.
$$

This implies that

$$\frac{12}{b^2}\mathbb{E}\sum_{j\in J_k}\left\|\nabla f_j(\varphi_k^j)-\nabla f_j(\xi_k^j)\right\|^2 \le \frac{12}{bn^2}\sum_{\ell=1}^{k}\left(1-\frac{b}{n}\right)^{k-\ell}\sum_{i=1}^{n}\|\nabla f_i(x_{\ell+1})-\nabla f_i(x_\ell)\|^2$$
$$\le \frac{12}{bn^2}\sum_{\ell=1}^{k}\left(1-\frac{b}{2n}\right)^{k-\ell}\sum_{i=1}^{n}\|\nabla f_i(x_{\ell+1})-\nabla f_i(x_\ell)\|^2.$$

We include the inequality of the second line to simplify later arguments. This completes our bound for the first term of (15). For the second term, we recall Lemma 21.

$$\mathbb{E}\left\|\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}}-\nabla f(x_k)\right\|^2 \le \frac{3}{b^2}\sum_{\ell=1}^{k}\left(1-\frac{b}{2n}\right)^{k-\ell}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_\ell)-\nabla f_i(x_{\ell-1})\|^2.$$

Combining all of these bounds, we have shown

$$\begin{aligned}&\mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARGE}}-\widetilde{\nabla}_k^{\text{SARGE}}\|^2\\ \le&\frac{12}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2+\frac{27b^2+12b}{n^2b^2}\sum_{\ell=2}^{k}\left(1-\frac{b}{2n}\right)^{k-\ell}\sum_{i=1}^{n}\|\nabla f_i(x_{\ell-1})-\nabla f_i(x_{\ell-2})\|^2\\ &+\frac{3b^2}{2n^2}\left\|\nabla f(x_k)-\widetilde{\nabla}_k^{\text{SARGE}}\right\|^2.\end{aligned}$$

□

Lemma 22 allows us to take advantage of the recursive structure of our gradient estimate. With this lemma established, we can prove a bound on the MSE.

Lemma 23 *The SARGE gradient estimator satisfies the following recursive bound:*

$$\begin{aligned}&\mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARGE}}-\nabla f(x_{k+1})\|^2\\ \le&\left(1-\frac{b}{n}+\frac{3b^2}{2n^2}\right)\mathbb{E}\left\|\widetilde{\nabla}_k^{\text{SARGE}}-\nabla f(x_k)\right\|^2+\frac{12}{bn}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\|^2\\ &+\frac{27b^2+12b}{n^2b^2}\sum_{\ell=1}^{k}\left(1-\frac{b}{2n}\right)^{k-\ell}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(x_\ell)-\nabla f_i(x_{\ell-1})\|^2.\end{aligned}$$

Proof. The beginning of our proof is similar to the proof of the variance bound for the SARAH gradient estimator in [23, Lem. 2].

$$\begin{aligned}&\mathbb{E}_k\|\widetilde{\nabla}_{k+1}^{\text{SARGE}}-\nabla f(x_{k+1})\|^2\\ =&\mathbb{E}_k\left\|\widetilde{\nabla}_k^{\text{SARGE}}-\nabla f(x_k)+\nabla f(x_k)-\nabla f(x_{k+1})+\widetilde{\nabla}_{k+1}^{\text{SARGE}}-\widetilde{\nabla}_k^{\text{SARGE}}\right\|^2\\ =&\left\|\widetilde{\nabla}_k^{\text{SARGE}}-\nabla f(x_k)\right\|^2+\|\nabla f(x_k)-\nabla f(x_{k+1})\|^2+\mathbb{E}_k\left\|\widetilde{\nabla}_{k+1}^{\text{SARGE}}-\widetilde{\nabla}_k^{\text{SARGE}}\right\|^2\\ &+2\langle\nabla f(x_k)-\widetilde{\nabla}_k^{\text{SARGE}},\nabla f(x_{k+1})-\nabla f(x_k)\rangle\\ &-2\left\langle\nabla f(x_k)-\widetilde{\nabla}_k^{\text{SARGE}},\mathbb{E}_k\left[\widetilde{\nabla}_{k+1}^{\text{SARGE}}-\widetilde{\nabla}_k^{\text{SARGE}}\right]\right\rangle\\ &-2\left\langle\nabla f(x_{k+1})-\nabla f(x_k),\mathbb{E}_k\left[\widetilde{\nabla}_{k+1}^{\text{SARGE}}-\widetilde{\nabla}_k^{\text{SARGE}}\right]\right\rangle.\end{aligned}$$

We consider each inner product separately. The first inner product is equal to

$$\begin{aligned}2\langle\nabla f(x_k)-\widetilde{\nabla}_k^{\text{SARGE}},\nabla f(x_{k+1})-\nabla f(x_k)\rangle =&-\|\nabla f(x_k)-\widetilde{\nabla}_k^{\text{SARGE}}\|^2-\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2\\ &+\|\nabla f(x_{k+1})-\widetilde{\nabla}_k^{\text{SARGE}}\|^2.\end{aligned}$$

For the next two inner products, we use the fact that

$$\begin{aligned}
&\mathbb{E}_k\left[\widetilde{\nabla}_{k+1}^{\text{SARGE}} - \widetilde{\nabla}_k^{\text{SARGE}}\right] \\
=&\mathbb{E}_k\left[\widetilde{\nabla}_{k+1}^{\text{SAGA}} - \left(1-\frac{b}{n}\right)\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}} + \left(1-\frac{b}{n}\right)\widetilde{\nabla}_k^{\text{SARGE}}\right] - \widetilde{\nabla}_k^{\text{SARGE}} \\
=&\nabla f(x_{k+1}) - \left(1-\frac{b}{n}\right)\nabla f(x_k) - \frac{b}{n}\widetilde{\nabla}_k^{\text{SARGE}} \\
=&\nabla f(x_{k+1}) - \nabla f(x_k) + \frac{b}{n}\left(\nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}}\right).
\end{aligned}$$

With this equality established, we see that the second inner product is equal to

$$\begin{aligned}
&-2\left\langle \nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}}, \mathbb{E}_k\left[\widetilde{\nabla}_{k+1}^{\text{SARGE}} - \widetilde{\nabla}_k^{\text{SARGE}}\right]\right\rangle \\
=&-2\langle \nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}}, \nabla f(x_{k+1}) - \nabla f(x_k)\rangle - \frac{2b}{n}\langle \nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}}, \nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}}\rangle \\
=&\|\nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}}\|^2 + \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 - \|\nabla f(x_{k+1}) - \widetilde{\nabla}_k^{\text{SARGE}}\|^2 - \frac{2b}{n}\|\nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}}\|^2 \\
=&\left(1-\frac{2b}{n}\right)\|\nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}}\|^2 + \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 - \|\nabla f(x_{k+1}) - \widetilde{\nabla}_k^{\text{SARGE}}\|^2.
\end{aligned}$$

The third inner product can be bounded using a similar procedure.

$$\begin{aligned}
&-2\left\langle \nabla f(x_{k+1}) - \nabla f(x_k), \mathbb{E}_k\left[\widetilde{\nabla}_{k+1}^{\text{SARGE}} - \widetilde{\nabla}_k^{\text{SARGE}}\right]\right\rangle \\
=&-2\langle \nabla f(x_{k+1}) - \nabla f(x_k), \nabla f(x_{k+1}) - \nabla f(x_k)\rangle - \frac{2b}{n}\langle \nabla f(x_{k+1}) - \nabla f(x_k), \nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}}\rangle \\
\leq&-2\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 + \frac{b}{n}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 + \frac{b}{n}\|\nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}}\|^2 \\
=&-\left(2-\frac{b}{n}\right)\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 + \frac{1}{n}\|\nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}}\|^2,
\end{aligned}$$

where the inequality is Young's. Altogether and after applying the full expectation operator, we have

$$\begin{aligned}
&\mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARGE}} - \nabla f(x_{k+1})\|^2 \\
\leq&\left(1-\frac{b}{n}\right)\mathbb{E}\left\|\widetilde{\nabla}_k^{\text{SARGE}} - \nabla f(x_k)\right\|^2 - \left(1-\frac{b}{n}\right)\mathbb{E}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 + \mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARGE}} - \widetilde{\nabla}_k^{\text{SARGE}}\|^2 \\
\leq&\left(1-\frac{b}{n}\right)\mathbb{E}\left\|\widetilde{\nabla}_k^{\text{SARGE}} - \nabla f(x_k)\right\|^2 + \mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARGE}} - \widetilde{\nabla}_k^{\text{SARGE}}\|^2.
\end{aligned}$$

Finally, we bound the last term on the right using Lemma 22.

$$\begin{aligned}
&\mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARGE}} - \nabla f(x_{k+1})\|^2 \\
\leq&\left(1-\frac{b}{n}+\frac{3b^2}{2n^2}\right)\mathbb{E}\left\|\widetilde{\nabla}_k^{\text{SARGE}} - \nabla f(x_k)\right\|^2 + \frac{12}{bn}\sum_{i=1}^n \mathbb{E}\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|^2 \\
&+\frac{27b^2+12b}{n^2b^2}\sum_{\ell=1}^k\left(1-\frac{b}{2n}\right)^{k-\ell}\sum_{i=1}^n \mathbb{E}\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2.
\end{aligned}$$

□

Lemma 23 shows that the SARGE gradient estimator satisfies the MSEB property with suitably chosen parameters.

Corollary 24 *The SARGE gradient estimator with $b \leq n/3$ satisfies the MSEB property with $M_1 = 12/b$, $M_2 = \frac{27b^2+12b}{n^2b^2}$, $\rho_M = \frac{b}{2n}$, $\rho_B = b/n$, and $\rho_F = \frac{b}{2n}$.*

Proof. It is easy to see that $\rho_B = b/n$ by computing the expectation of the SARGE gradient estimator.

$$\begin{aligned}\nabla f(x_{k+1}) - \mathbb{E}_k \widetilde{\nabla}_{k+1}^{\text{SARGE}} &= \nabla f(x_{k+1}) - \mathbb{E}_k \left[\widetilde{\nabla}_{k+1}^{\text{SAGA}} - \left(1 - \frac{b}{n}\right) \left(\widetilde{\nabla}_{k+1}^{\xi\text{-SAGA}} - \widetilde{\nabla}_k^{\text{SARGE}} \right) \right] \\ &= \left(1 - \frac{b}{n}\right) \left(\nabla f(x_k) - \widetilde{\nabla}_k^{\text{SARGE}} \right).\end{aligned}$$

The result of Lemma 23 makes it clear that $M_1 = 12/b$. To determine ρ_M, we must first choose a suitable sequence $\mathcal{M}_k$. Let $\mathcal{M}_k = \mathbb{E}\|\widetilde{\nabla}_{k+1}^{\text{SARGE}} - \nabla f(x_{k+1})\|^2$. If $n = 1$, then $\mathcal{M}_k = 0$ for all k, so it holds trivially that $\mathcal{M}_k \leq (1 - \rho_M)\mathcal{M}_{k-1}$. If $n \geq 2$, the fact that $b \leq n/3$ ensures that $1 - \frac{b}{n} + \frac{3b^2}{2n^2} \leq 1 - \frac{b}{2n}$, so Lemma 23 ensures that with $\rho_M = \frac{b}{2n}$, $\mathcal{M}_k \leq (1 - \rho_M)\mathcal{M}_{k-1}$.

Finally, we must compute M_2 and ρ_F with respect to some sequence $\mathcal{F}_k$. Lemma 23 motivates the choice

$$\mathcal{F}_k = \sum_{\ell=1}^{k} \left(1 - \frac{b}{2n}\right)^{k-\ell} \sum_{i=1}^{n} \mathbb{E}\|\nabla f_i(x_\ell) - \nabla f_i(x_{\ell-1})\|^2,$$

and the choices $M_2 = \frac{27+12b}{n^2b^2}$ and $\rho_F = \frac{b}{2n}$ are clear. □

To prove the convergence rates of Theorem 16, we simply combine the MSEB constants of Corollary 24 with Theorems 5 and 6.